Hanuman at Touché: Image Generation with Argument-Aspect Fusion

Notebook for the Touché Lab at CLEF 2025

Sharat Anand[1], Maximilian Heinrich[1]

[1]Bauhaus-Universität, Weimar, Germany

Abstract

Generating images from textual arguments presents a significant challenge due to the inherent ambiguity and context-dependence of natural language. For the Touché 2025 shared task 'Image Retrieval/Generation for Arguments', we present a system that generates an image for an argument by first identifying and extracting key aspects of the argument. These aspects are then combined with the original argument to form a detailed prompt that describes a situation relevant to the argument. Finally, this prompt is used to generate a corresponding image. This approach allows us to guide image generation more effectively, ensuring that the output more accurately reflects the intended meaning of the argument. To evaluate the results, we employ several embedding methods to measure the semantic similarity between the arguments and the corresponding images. Our findings indicate that incorporating dedicated aspects into the image prompts significantly improves the quality and relevance of the generated images.

Keywords

Image Generation for Arguments, Argument Aspect Extraction, Multimodal Argument Generation

1. Introduction

Images are a powerful tool for supporting arguments and revealing the underlying components of complex ideas [1]. Their unique ability to evoke emotion and highlight critical aspects of a message underscores the potential of integrating visual elements into argumentative discourse. The Touché 2025 shared task on 'Image Retrieval/Generation for Arguments' [2, 3] focuses on developing automated methods to align arguments with images, challenging participants to retrieve or generate images that effectively match the arguments. In this work, we present an argument-image generation system developed for the shared task. Our approach focuses on identifying the core aspects of an argument and using them to construct a prompt that emphasizes these key elements. We then assess whether the generated images effectively capture the intended argumentative content. The remainder of this paper is structured as follows: we begin with an overview of the shared task and related work, followed by a detailed description of our system. Finally, we present an evaluation of the generated images, focusing on their quality and relevance in representing the extracted argumentative aspects.

2. Task Description

The objective of the shared task is to retrieve or generate images that effectively convey the central claims of given arguments. Participants can either select images from a provided dataset of approximately 32,000 web-crawled images or employ an image generation model of their choice. The task comprises 128 arguments covering 27 distinct topics [2]. These arguments are typically brief — for example, "Hiker's trash contributes to environmental damage." For each argument, participants are required to submit five ranked images via the TIRA shared task platform [4].

Identifying suitable images for textual arguments remains a significant challenge, as images are "ambiguous, yet rich in information" [5]. While images often convey more than words alone, their

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

stsharatanand@gmail.com (S. Anand); maximilian.heinrich@uni-weimar.de (M. Heinrich)

https://webis.de/publications.html (S. Anand)

interpretation frequently relies on contextual cues. In previous iterations of the shared task, participants have explored how image generation can enhance retrieval — for instance, by producing reference images to compare against existing ones within a retrieval-based framework [6, 7]. In contrast, our approach is designed exclusively for image generation, without incorporating any retrieval component.

3. System Description

Initial experiments using the raw argument text as a prompt for image generation revealed a recurring issue: many of the resulting images failed to capture core aspects of the input arguments. Instead of conveying the intended semantic content, the generated images often included irrelevant or distracting visual elements, introducing noise that compromised both their interpretability and overall effectiveness. To address this limitation, our approach focuses on visualizing the key aspects of each argument. This is achieved through the integration of several dedicated modules. An overview of our system architecture is shown in Figure 1.

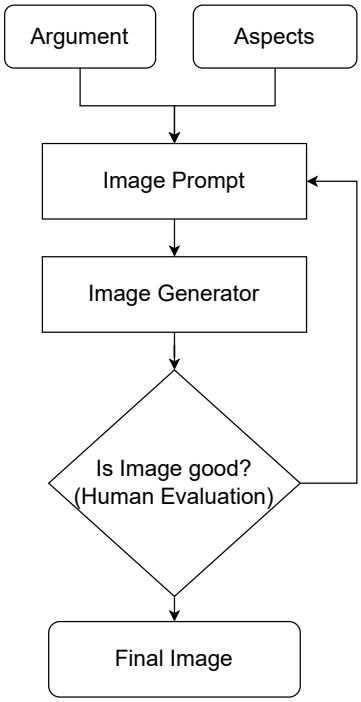
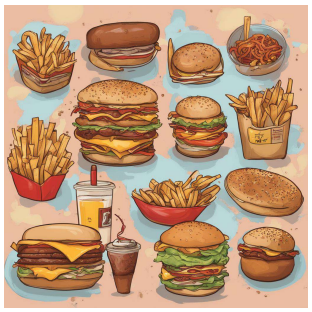


Figure 1: Overview of the argument-to-image generation pipeline. After identifying the relevant aspects of the argument, both the argument and its extracted aspects are fed into the image prompt generator, which produces a detailed prompt. This prompt is then passed to an image generation model to create the corresponding image. The generated image is subsequently evaluated for semantic relevance and visual quality. If the image does not meet the evaluation criteria, feedback is used to iteratively refine the prompt until a satisfactory image is achieved.

The first step in our system is to identify which aspects of the argument should be visualized in the corresponding image. To do this, we use the LLaMA 3.2 (3B-Instruct) language model [8, 9] to extract three key aspects that capture the most important elements of the argument. These aspects serve as essential components that must be visually represented in the generated image. In the second step, the list of extracted aspects, together with the original argument, is passed to the argument-image prompt generator module. This module leverages the Mistral (7B-v0.1) model [10, 11] to create a detailed prompt for the image generator (image prompt). Mistral was chosen for this task due to its superior performance compared to other large language models [10]. The image prompt describes the argument as a vivid scene, carefully incorporating each identified aspect. Table 1 demonstrates how the final image prompt differs from the original argument. In the third step, the image prompt is utilized to generate

Table 1
Comparison of prompting methods using Mistral (7B-v0.1). This comparison evaluates the difference between using a simple argument directly as a prompt versus employing a more detailed prompt that integrates key aspects of the image (image prompt). These key aspects are automatically extracted using LLaMA 3.2 (3B-Instruct). As demonstrated, incorporating relevant aspects into the prompt substantially improves the visual clarity and overall quality of the generated image.

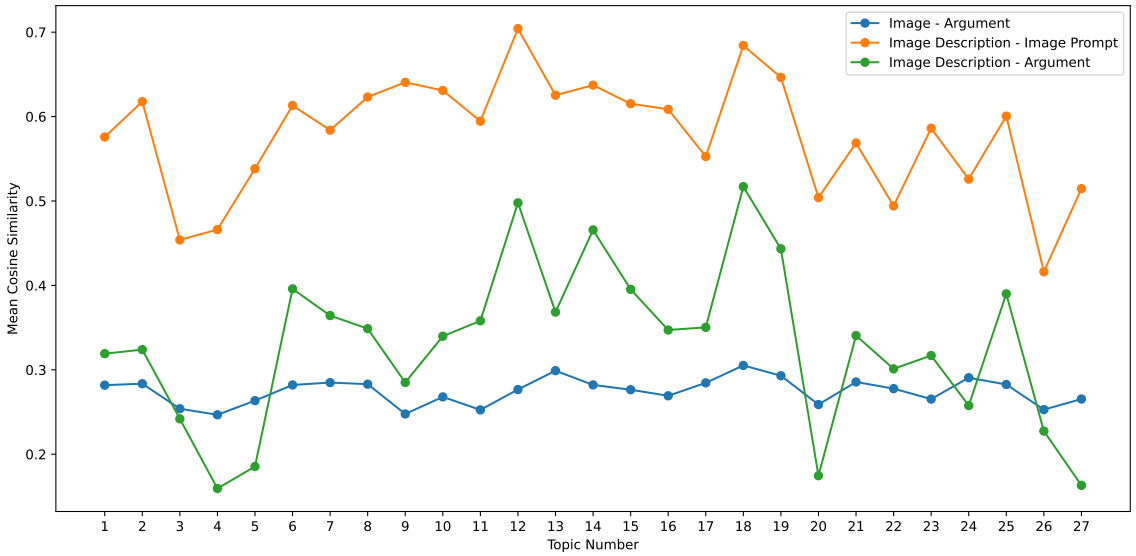
Argument: Consuming too much fast food leads to obesity.	
Identified Aspects: Fast-food Consumption, Obesity, Overeating **Image Prompt:** A man sitting in a fast food restaurant, surrounded by empty wrappers and packaging. He has a large belly, and his face is red and sweaty from overeating.	

the corresponding image. We employ Stable Diffusion XL (Base 1.0) [12, 13] as the image generation model, configured with a fixed seed value of 1244 and 40 inference steps to ensure consistency and reproducibility. This model is selected over alternative models due to its capability for producing softer, more artistic renderings — particularly well-suited for stylized, anime, and fantasy art — as evidenced by the comparative results presented in Table 1. The fourth step involves a quality check performed by a human expert, who evaluates whether the generated image effectively represents the key aspects of the argument. If the image is deemed satisfactory, it is accepted; otherwise, the prompt is revised to emphasize any missing elements. However, for the images submitted to the task, no prompt revisions were required. Using this process, five images were generated for each argument.

Finally, the generated images are ranked by first generating a description of each image using LLaVA 1.5 (13B) [14, 15], and then computing the cosine similarity between this description and the original image prompt using Sentence-BERT (SBERT) [16], specifically the all-MiniLM-L6-v2 model. An example of this similarity assessment and ranking method is illustrated in Table 2 in the appendix.

4. Evaluation

To evaluate how well the generated images align with the arguments, we employ three distinct methods to measure semantic similarity between argument-image pairs. The first method directly compares the image and the argument by computing cosine similarity between their embeddings, generated using the multimodal CLIP model (clip-vit-base-patch32) [17]. For the second and third methods, we first generate a description of each image using LLaVA 1.5 (13B) [14, 15]. In the second method, the image description is compared to the corresponding argument by computing SBERT embeddings and measuring their cosine similarity. The third method compares the image description to the image prompt used to generate the image, also using SBERT embeddings. This third method is also employed internally by the system to determine the final ranking of the generated images, as detailed in Section 3. A summary of the mean cosine similarity scores for all three methods across individual topics, along



Evaluation	Mean	Median	Std. Dev.	Min	Max
Image - Argument	0.2757	0.2779	0.0310	0.1613	0.3645
Image Description - Image Prompt	0.5719	0.5971	0.1626	0.0892	0.8972
Image Description - Argument	0.3266	0.3288	0.1483	0.0033	0.6707

Figure 2: Metrics of cosine similarity for various embedding methods. (Top:) Variation in mean cosine similarity across 27 topics using three different embedding methods. The blue line represents CLIP embeddings between the image and the argument. The orange and green lines represent SBERT embeddings computed between the automatically generated image description (produced by LLaVA) and the image prompt used to generate the image (orange) or the argument (green). (Bottom:) Averaged similarity scores across all topics, showing mean, median, standard deviation (Std. Dev.), minimum, and maximum cosine similarity values.

with central metrics averaged across all approaches, is presented in Figure 2. More detailed results for each of the three evaluation methods are available in the appendix, as shown in Figure 3. The results show that the cosine similarity between the image description generated by LLaVA and the image prompt is the highest among all measured similarities, consistently exceeding the similarity between the image description and the original argument. This is likely because the image prompts contain more detailed descriptions of visual aspects than the arguments alone, leading to higher similarity scores. Overall, the consistently high similarity confirms that the generated images closely align with the image prompts. In addition, we observe a strong correlation between similarities based on the image prompt and those based on the argument. This indicates that the image prompt is thematically aligned with the argument. Some topics in Figure 2 exhibit noticeably lower similarity. These cases may correspond to complex arguments, or to arguments that require the visualization of undesirable or negative aspects.

The CLIP-based similarity between the image and the original argument consistently produces the lowest similarity scores, with the values lying very close together. This observation aligns with existing research [18], which highlights that CLIP similarities are generally very uniform. This insight plays a particularly important role in automatically deciding whether an image fits an argument. In such cases, CLIP embeddings are likely not suitable.

Notably, as shown in Table 2 in the appendix, the image prompts used to guide generation differ substantially from the original arguments. Through this transformation and fusion, the visual essence of the argument is better captured, resulting in improved image quality, enhanced visual clarity, and reduced visual noise. The strength of our approach for generating highly relevant images is demonstrated by its first-place ranking in the shared task, outperforming all baseline approaches that rely on the raw argument text as prompts.

5. Conclusion

In this work, we introduced a novel argument-image generation system that first identifies relevant aspects of the arguments and then combines these aspects with the original argument to create a corresponding image prompt. This prompt is used by an image generator to produce the related image. By enriching the original arguments with key aspects and using detailed, extended prompts for generation, our approach improves image quality and reduces visual noise. This enables the generation of highly relevant images and resulted in the system achieving first place in the competition.

Acknowledgments

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

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, DeepL, Grammarly, Grok and Language Tool in order to: grammar and spelling check, paraphrase and reword, improve writing style. The presented image generation pipeline integrates several AI models: prompts are produced with LLaMA, Stable Diffusion is used to generate images, and LLaVA provides automated image descriptions. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. J. Dove, On images as evidence and arguments, in: F. H. van Eemeren, B. Garssen (Eds.), Topical Themes in Argumentation Theory: Twenty Exploratory Studies, Argumentation Library, Springer Netherlands, Dordrecht, 2012, pp. 223–238. doi:10.1007/978-94-007-4041-9_15.

[2] J. Kiesel, Ç. Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation Systems, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[3] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[4] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances

in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[5] J. E. Kjeldsen, The rhetoric of thick representation: How pictures render the importance and strength of an argument salient, Argumentation 29 (2015) 197–215. URL: https://doi.org/10.1007/s10503-014-9342-2. doi:10.1007/s10503-014-9342-2.

[6] B. Ostrower, P. Aphiwetsa, Ds@gt at touché: Image search and ranking via clip and image generation, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 3447–3450. URL: http://ceur-ws.org/Vol-3740/paper-332.pdf.

[7] M. Moebius, M. Enderling, S. T. Bachinger, Jean-luc picard at touché 2023: Comparing image generation, stance detection and feature matching for image retrieval, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes Papers of the CLEF 2023 Evaluation Labs, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 3111–3118. URL: http://ceur-ws.org/Vol-3497/paper-263.pdf.

[8] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, et al., The llama 3 herd of models, CoRR abs/2407.21783 (2024). URL: https://doi.org/10.48550/arXiv.2407.21783. doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

[9] Meta AI, meta-llama/llama-3.2-3b-instruct model card, https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024. Model card accessed in June 2025.

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, CoRR abs/2310.06825 (2023). URL: https://doi.org/10.48550/arXiv.2310.06825. doi:10.48550/ARXIV.2310.06825. arXiv:2310.06825.

[11] Mistral AI Team, mistralai/mistral-7b-v0.1 model card, https://huggingface.co/mistralai/Mistral-7B-v0.1, 2023. Hugging Face model card accessed June 2025.

[12] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: improving latent diffusion models for high-resolution image synthesis, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024. URL: https://openreview.net/forum?id=di52zR8xgf.

[13] Stability AI, stabilityai/stable-diffusion-xl-base-1.0 Model Card, https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0, 2023. Model card accessed June 2025; includes details on ensemble pipeline, text encoders, and license (OpenRAIL++-M).

[14] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), 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, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

[15] LLaVA Contributors, user/llava-1.5-13b-hf model card, https://huggingface.co/user/llava-1.5-13b-hf, 2025. Hugging Face model card accessed June 2025; variant of the LLaVA-1.5 13 B model with HF integration.

[16] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks,

in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.

[17] OpenAI, openai/clip-vit-base-patch32 model card, https://huggingface.co/openai/clip-vit-base-patch32, 2021. Model card accessed June 2025, covers architecture, intended use, performance, and biases.

[18] K. Tyshchuk, P. Karpikova, A. Spiridonov, A. Prutianova, A. Razzhigaev, A. Panchenko, On isotropy of multimodal embeddings, Information-an International Interdisciplinary Journal 14 (2023). doi:10.3390/info14070392.

Appendix

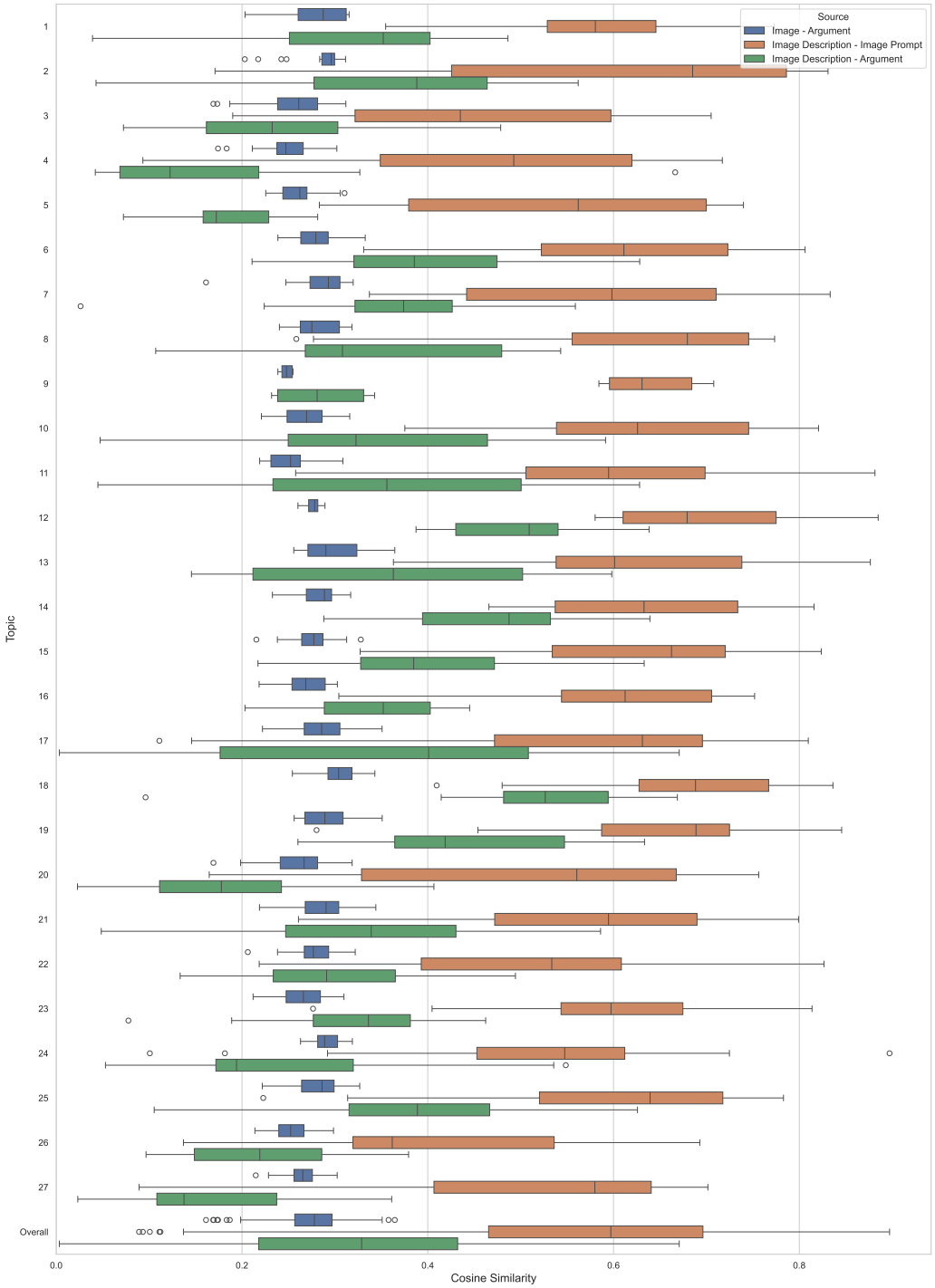


Figure 3: Similarity Distribution. The boxplot illustrates the distribution of cosine similarity scores across 27 topics using three different embedding methods. The blue boxes represent CLIP embeddings computed between the image and the argument. The orange and green boxes represent Sentence-BERT embeddings between the automatically generated image description (produced by LLaVA) and either the image prompt used to generate the image (orange) or the argument (green).

Table 2
Comparison between arguments and the image prompts used to generate the corresponding images. The first column contains the original argument from the dataset. The 'Image Prompt' column displays the prompt generated by the language model, incorporating key aspects of the argument to guide image generation. The 'Image Description' column shows the image-to-text description produced by the LLaVA model. The 'CS' (Cosine Similarity) column reports the similarity between the generated prompt and the LLaVA description, computed using the SBERT model. The final column provides the ranking based on cosine similarity scores submitted for the shared task.

Argument	Image Prompt	Image Description	CS	Rank
Consuming too much Fast Food leads to obesity	A young adult woman is sitting at a table with a plate of french fries, a burger, and a soda in front of her. She is looking down at the food and is frowning, clearly feeling guilty about her unhealthy eating habits. She is wearing jeans and a t-shirt, and her hair is pulled back in a ponytail. Behind her, there is a McDonald's sign and a trash can overflowing with fast food wrappers.	In the image, a woman is sitting at a dining table with a plate of food in front of her. The plate contains a hamburger, french fries, and a sandwich. She is holding a sandwich in her hand, and there is a cup on the table as well. The woman appears to be enjoying her meal, and the scene is set in a restaurant.	0.79	1
Consuming too much Fast Food leads to obesity	A man sitting in a fast food restaurant, surrounded by empty wrappers and packaging. He has a large belly, and his face is red and sweaty from overeating.	The image is a cartoon drawing of a man sitting on a chair. He is wearing a white shirt and red pants. The man appears to be quite large, possibly obese. He is surrounded by a large number of bags of chips, with some of them scattered on the floor. The man seems to be enjoying his time, possibly eating the chips.	0.74	2
Consuming too much Fast Food leads to obesity	A group of children playing in a park, surrounded by fast food containers, chips, and sugary drinks. The children are oblivious to the unhealthy consequences of their snacking habits, and continue to eat unhealthy food, ignoring the warnings from their parents and health experts.	The image depicts a group of children gathered around a tree, enjoying a picnic together. They are sitting on the ground, with some of them eating chips and drinking from cups. There are several cups scattered around the area, along with a few bottles. The children are engaged in conversation and laughter, creating a lively and fun atmosphere. The scene captures the essence of childhood and the simple joys of spending time outdoors with friends.	0.54	3