Understanding Modality Preferences in Search Clarification

Leila Tavakoli*, Giovanni Castiglia, Federica Calò, Yashar Deldjoo, Hamed Zamani and Johanne R. Trippas

RMIT University, Australia
Polytechnic University of Bari, Italy
Polytechnic University of Bari, Italy
Polytechnic University of Bari, Italy
University of Massachusetts Amherst, United States
RMIT University, Australia

Abstract

This study is the first attempt to explore the impact of clarification question modality on user preference in search engines. We introduce the *multi-modal* search clarification dataset, *MIMICS-MM*, containing clarification questions with associated expert-collected and model-generated images. We analyse user preferences over different clarification modes of text, image, and combination of both through crowdsourcing by taking into account image and text quality, clarity, and relevance. Our findings demonstrate that users generally prefer *multi-modal* clarification over *uni-modal* approaches. We explore the use of automated image generation techniques and compare the quality, relevance, and user preference of model-generated images with human-collected ones. The study reveals that text-to-image generation models, such as *Stable Diffusion*, can effectively generate *multi-modal* clarification questions. By investigating *multi-modal* clarification, this research establishes a foundation for future advancements in search systems.

Keywords

multi-modal clarification, search clarification dataset, text-to-image generation,

1. Introduction

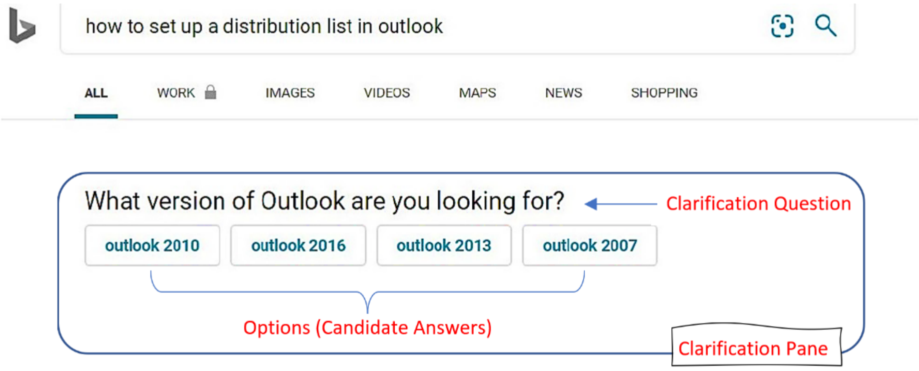
Effective communication between users and intelligent systems is essential for accurately identifying a user's information needs. One common obstacle encountered by Information Retrieval systems is the inherent ambiguity present in human language. Clarification questions can play a pivotal role in search interactions, allowing users to refine their queries and obtain more precise search results. Traditionally, clarification questions have been presented in a textual format, allowing users to respond with further textual input. Figure 1 shows an example of a clarification pane presented to users on the *Bing* search engine. In this scenario, the user is seeking information about setting up a distribution list in Outlook, and the clarification question aims to clarify the version of Outlook that the user is working with. While clarification has become an important component of many conversational and interactive information-seeking systems [1], previous research has shown that even though clarification questions receive positive engagement, users are not frequently engaged with them [2, 3].

Recent advancements in technology have introduced new modalities, such as visual prompts or multi-modal, which is a combination of text and visuals. As emphasised in the most recent Alexa Prize TaskBot Challenge [4], there are instances in which *multi-modal* interactions (e.g., text and image) impact the user experience in conversational information-seeking systems [5].

1st Workshop on Multimodal Search and Recommendations (CIKM MMSR '24), October 25 2024, Boise, Idaho, USA

*Corresponding author.

leila.tavakoli31@gmail.com (L. Tavakoli); g.castiglia@studenti.poliba.it (G. Castiglia); f.calo8@studenti.poliba.it (F. Calò); yashar.deldjoo@poliba.it (Y. Deldjoo); zamani@cs.umass.edu (H. Zamani); j.trippas@rmit.edu.au (J. R. Trippas)
0000-0002-5951-4052 (L. Tavakoli); 0000-0002-6767-358X (Y. Deldjoo); 0000-0002-0800-3340 (H. Zamani); 0000-0002-7801-0239 (J. R. Trippas)



how to set up a distribution list in outlook

What version of Outlook are you looking for?

Figure 1: A clarification pane shown after a user query [3].

Although incorporating visual elements allows users to provide more context and improve query accuracy, the extent to which different modalities (such as text or image) can enhance user interaction in search engines is still uncertain. Previous studies have primarily focused on text-based clarification, neglecting the potential benefits of *multi-modal* approaches. By exploring user preferences over various modalities, we can investigate which modalities are perceived as more effective and intuitive for optimising both user experience and system performance. We study *multi-modal* clarification questions from the user behaviour perspective and explore user preference on clarification question modalities, specifically focusing on *text-only*, *visual-only*, and *multi-modal* approaches. By systematically analysing user feedback, we can gain valuable insights into the advantages and limitations of each modality and the influential parameters.

A clarification pane typically consists of a multi-choice clarification question and a list of candidate answers [3]. A *multi-modal* clarification pane contains both visual and textual content for each candidate answer (see Figure 2). We aim to understand if adding a visual presentation to *text-only* clarification panes enhances the user experience.

We explore user preferences over three modalities for clarification panes: *(i)* textual, *(ii)* visual, and *(iii) multi-modal* (i.e., a combination of the two). We randomly sample 100 query-clarification pairs from the *MIMICS* dataset [3]. Then, we create the visual and *multi-modal* clarification panes for the sampled query-clarification pairs through a controlled manual expert annotation process. Pairwise user preferences are collected for different modalities following a post-task questionnaire to answer the following research question:

- Do users prefer *multi-modal* clarification panes over *uni-modal* (i.e., textual or visual)?

In this study, we investigate the impact of the image quality, image/text clarity, and relevance of the text and image, in addition to various image aspects, on user preference. Finally, we explore whether generating corresponding images to the clarification panes can be automated using text-to-image generation models. The quality and the relevance of generated images, in addition to user preferences over human-collected and model-generated images, are investigated through manual annotation. Our experiments reveal that:

- In the majority of cases (70-80%), users prefer *multi-modal* clarification panes over *visual-only* and *text-only* clarification panes. They also prefer *visual-only* clarification over *text-only* clarification in 54% of cases.

- Crowd-source workers prefer *multi-modal* clarification panes as they are easier to understand, which helps users make better and faster decisions.

- Image quality, clarity, and relevance, in addition to text clarity, have a direct impact on self-reported user perceptions.

- Text-to-image generation models, such as Stable Diffusion [6], are capable of automating image generation for creating *multi-modal* clarification panes.

Our contributions to this paper include:

- Gaining a better understanding of user preferences when it comes to different clarification modalities.
- Evaluating the influence of image and text properties on user preference. By investigating how different factors related to images and text affect user choices, we gain insights into the impact of these properties on search clarification.
- Exploring the capabilities of text-to-image generation models in the context of search clarification. By studying the effectiveness of these models in generating relevant images based on textual queries, we investigate their potential use in enhancing the search clarification process.

Overall, our findings provide valuable insights into how to engage the user better with clarifications in information-seeking systems. By understanding user preferences and leveraging *multi-modal* approaches, we can create more effective systems that cater to the needs of users in search clarification scenarios.

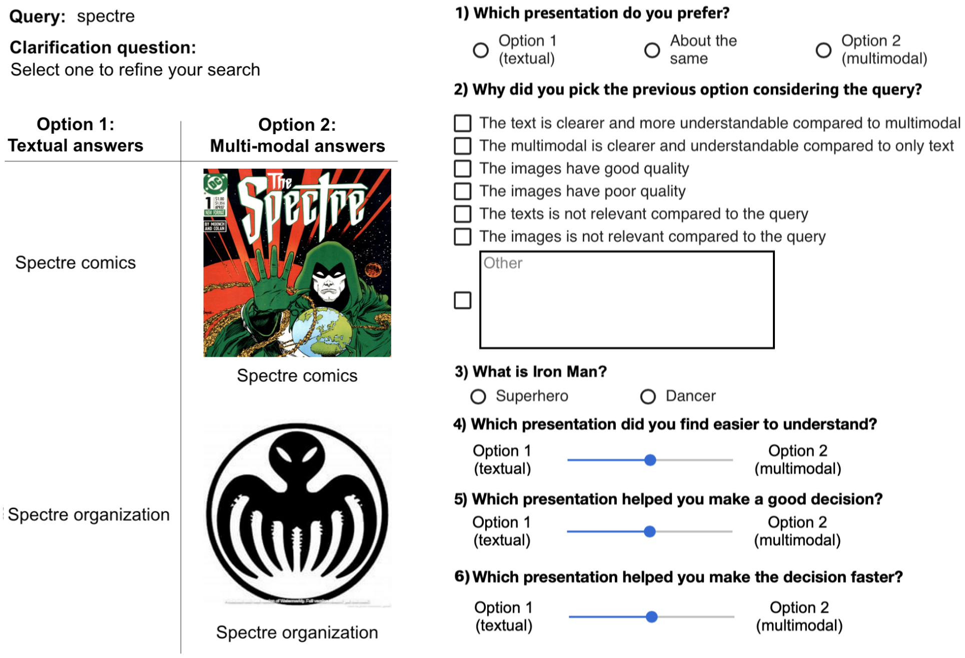
2. Related Work

Despite the growing interest in search clarification [7, 8, 9, 10, 11], there is a need for more research on improving user interaction with clarification questions. In addition, further integration of these approaches with the latest developments in multi-modal generative AI and IR systems could lead to more effective and intuitive user experiences [12]. Previous researchers such as Rao and Daumé III [13], Aliannejadi et al. [9], Zamani et al. [2], Sekulić et al. [14] primarily focused on enhancing the effectiveness of clarification modals in search systems. Still, there is a research gap regarding user preferences and perceptions of different modalities in search clarification. This literature review highlights that research on multi-modal IR has overlooked search clarifications. For example, Yang et al. [15] introduced an online video recommendation system incorporating multi-modal fusion and relevance feedback. While Zha et al. [16] proposed Visual Query Suggestion for image search, Altinkaya and Smeulders [17] developed a stuttering detection model, Srinivasan and Setlur [18] explored utterance recommendations for visual analysis, Pantazopoulos et al. [19] integrated computer vision and conversational systems for socially assistive robots, and Ferreira et al. [20] presented TWIZ, a multi-modal conversational task wizard. None of these works addressed multi-modal clarification questions in the context of search systems. Hence, this area has a significant research gap, highlighting the need for further exploration and development.

3. Experimental Design

We now describe the methodology and structure of the data collection, including the experiments.
Query and clarification panes sampling. We used the *MIMICS-Manual*[1] dataset to select textual clarification panes. We randomly selected 100 queries and their corresponding multi-choice clarification panes to create the *MIMICS-MM* dataset. The number of candidate answers in the clarification pane varies between two and five.
Clarification image collection. To assign an image to each candidate answer of clarification panes, an expert annotator searched the online website for corresponding images to those candidate answers

[1] The *MIMICS-Manual* is one of three subsets (i.e., *MIMICS-Click*, *MIMICS-ClickExplore*, and *MIMICS-Manual*) of the *MIMICS* dataset–the largest available search clarification dataset [3]. It contains over 2,000 search queries with multiple clarification panes, landing result pages and manually annotated three-point quality labels for clarification panes.



Query: spectre

Clarification question:
Select one to refine your search

Option 1: Textual answers	**Option 2: Multi-modal answers**
Spectre comics	Spectre comics
Spectre organization	Spectre organization

1) Which presentation do you prefer?

○ Option 1 (textual) ○ About the same ○ Option 2 (multimodal)

2) Why did you pick the previous option considering the query?

□ The text is clearer and more understandable compared to multimodal
□ The multimodal is clearer and understandable compared to only text
□ The images have good quality
□ The images have poor quality
□ The texts is not relevant compared to the query
□ The images is not relevant compared to the query
□ Other

3) What is Iron Man?

○ Superhero ○ Dancer

4) Which presentation did you find easier to understand?

Option 1 (textual) Option 2 (multimodal)

5) Which presentation helped you make a good decision?

Option 1 (textual) Option 2 (multimodal)

6) Which presentation helped you make the decision faster?

Option 1 (textual) Option 2 (multimodal)

Figure 2: An example of Task II (T vs. MM) with Post-task questionnaire (All of the questions were single-select).

using the Google images search engine.[2] In total, 314 images were matched with 314 textual candidate answers. The annotator re-evaluated the quality of the images and, if needed, replaced them with images of greater quality.

Experimental design. Online experiments[3] were conducted on Amazon Mechanical Turk (AMT) to gather user preference labels through Human Intelligence Tasks (HITs).[4] We designed three tasks to collect judgements from AMT workers on user preferences over different modalities in search clarification. We ran pairwise comparisons as follows:

- Task I: *text-only* (T) vs. *visual-only* (V)
- Task II: *text-only* (T) vs. *multi-modal* (MM)
- Task III: *visual-only* (V) vs. *multi-modal* (MM)

A query and two modalities are shown in Figure 2. At the end of this data collection process, three different subsets were created.

Post-task questionnaire. After showing a query and two clarification question options, workers were presented with a post-task questionnaire assessing their presentation style preference and feedback (Figure 2). Thus, after inspecting the query, clarification question, and candidate answers, workers indicated which presentation they preferred (*Q1*). Workers were also asked to justify their preference with four questions. The second question (*Q2*) contained checkboxes with options about the text and images' clarity, quality, and relevance. Workers were asked three more questions to obtain the motivation behind their choice of which modality was easier to understand (*Q4*), which helped them make better (*Q5*) and faster decisions (*Q6*) on a 5-point slider (e.g., in Task 2, labels 1 and 2 means *text-only* modality is preferred, label 0 means they have no preference, and labels 4, and 5 mean multi-modality is preferred).

[2] We watermarked the images for copyright compliance.

[3] Reviewed and approved according to RMIT University's ethics procedures for research involving human subjects. The approval number is 66-19/22334.

[4] Data collection was conducted in mid-March 2022.

Quality assurance. We included two quality assurance checks. For example, each task contained a gold question (i.e., a question with a known answer) with the aim of high validity throughout the task (see *Q3* in Figure 2). Workers who failed to answer the gold question were prohibited from completing other tasks, and their answers were removed. We also manually checked 10% of submitted HITs per task as a final quality assurance check. Invalid submissions were removed, and the workers were denied from completing subsequent tasks. We then opened those HITs to other workers. AMT pilot tasks were carried out[5] to analyse the flow, acquire users' feedback, check the quality of collected data, and estimate the required time to finish each task and a fair pay rate.

Workers. Only workers based in Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States, with a minimum HIT approval rate of 98% and a minimum of 5,000 accepted HITs, were allowed to participate in the study, maximising the collected data quality and the likelihood that workers were either native English speakers or had a high level of English. Each HIT was assigned to at least three different AMT workers, enabling us to use an agreement analysis measure on their modality preferences. The exact same number of users was assigned to each question to avoid creating bias towards giving more importance to the question with more assigned users. In case of disagreements, we administered the HIT again to more workers until we achieved a final majority vote. Each worker was allowed to perform 25 tasks (a portion used for each launch). Workers had a five-minute time limit to finish the task and were compensated with 0.74 USD per HIT.

Image generation for *multi-modal* clarification. Following a crowd-sourcing approach, we utilised two text-to-image generation models, namely Stable Diffusion[6] [6] and Dall·E 2[7] [21]. These models were employed to produce images related to candidate answers, with the aim of exploring their potential in generating multi-modal clarifications. Our input to generate a corresponding image to a candidate answer of a clarification pane was the concatenation of the query and the candidate answer text. This input was used to generate all corresponding images for all candidate answers (two employed models per candidate answer generated two images).

Comparing human-collected versus computer-generated images. First, we evaluated and compared the generated images' visual aspects with manually collected images. We extracted the visual aspects of the images using *OpenIMAJ* [22], a tool for multimedia content analysis. The nine visual aspects investigated were *brightness*, *colourfulness*, *naturalness*, *contrast*, *RGB contrast*, *sharpness*, *sharpness variation*, *saturation*, and *saturation variation* [23]. We conducted a manual annotation to investigate the generated images' relevance to the text, compare the images' quality, and assess the user preference over generated and collected images. Three annotators, two men and a woman with proficient English and a higher degree, completed the labelling. Each annotator labelled 314 generated images. We collected all annotations and aggregated them, and in case of any disagreements, majority voting was used for the final label.

We showed the concatenation of the query and the candidate answer in the text and the corresponding generated image to the annotators. We asked annotators if the image was relevant to the text or not on a binary scale (i.e., label 1 means relevant, and label 0 means irrelevant). This label was similar to the label collected for the human-collected images during crowd-sourcing. Then, we showed the collected image for the same text from the crowd-sourcing part and asked the annotators to compare the quality of generated and collected images regardless of the presented text on a 3-point scale (i.e., the quality of the computer-generated image is higher *(2)*, are the same *(1)*, or the human-collected image has a higher quality *(0)*). Finally, the annotators were asked to indicate their preferred image between two images on a 3-point scale (i.e., annotators prefer the computer-generated image *(2)*, have no preference *(1)*, or prefer the human-collected image *(0)*).

[5] Pilot study was conducted in February 2022.

[6] Stable Diffusion is a neural text-to-image model that uses a diffusion model variant called the latent diffusion model. It is capable of generating photo-realistic images given text input. The Diffusers library available at https://github.com/huggingface/diffusers is used for this study.

[7] Dall·E 2, created by OpenAI, generates synthetic images corresponding to an input text.

Table 1
Pairwise preference for clarification modality (%)

Task	Prefer Text	Prefer Visual	Prefer Multi-Modal	No preference
Text vs. Visual	$39^{\dagger}$	$54^{\dagger}$	NA	$7^{\dagger}$
Text vs. Multi-Modal	$17^{\dagger}$	NA	$79^{\dagger}$	$4^{\dagger}$
Visual vs. Multi-Modal	NA	17	$71^{\dagger}$	12

† Significantly different from the other two preferences (Tukey HSD test, *p<0.05*).

4. Results

In this section, we investigate the impact of various clarification modality characteristics and visual aspects of the images on user preference. Furthermore, we explore whether the clarification panes' visual modality can be automated.[8]

User preference and clarification modality. We first investigated user preferences over the clarification modality in each pairwise comparison (i.e., *text-only* vs. *visual-only*, *text-only* vs. *multi-modal*, and *visual-only* vs. *multi-modal*). To understand whether a preferred modality in each pairwise comparison is significantly different from the other two options, we performed the Tukey honestly significant difference (HSD)[9] test [25]. This statistical significance test helped us determine, for instance, if the number of users who preferred *multi-modal* over *text-only* was significantly higher or not. Table 1 indicates the percentage of user preference in each pairwise clarification modality comparison (i.e., The average across all the user inputs). In Task 1, where the workers indicated their preferences between the *text-only* and *visual-only* clarifications, we observed that 54% of the workers preferred *visual-only* over *text-only* clarification panes. In Tasks 2 and 3, where the workers indicated their preferences between *uni-modal* and *multi-modal* clarification panes, the workers strongly preferred *multi-modal* clarification panes, no matter whether the uni-modal clarification pane is *text-only* or *visual-only*. The workers' preferences were significantly different from other options, indicating that in 70-80% of the cases, a *multi-modal* clarification was preferred.

Post-task questionnaire analysis. We asked the workers to explain if the text/image clarity relevance and image quality impacted their preferences. We calculated the Pearson correlations between the workers' preferences and the characteristics of the clarification modalities in each Task. In Task 1, we observed a positive correlation (ρ=0.476) between user preference (i.e., preferring *visual-only* clarifications over *text-only* ones) and image quality. There was also a strong positive correlation (ρ=0.677) between user preference and image clarity, and user preference had a strong negative correlation (ρ=-0.686) with text clarity. The same correlation trends and orders were observed for the user preference (i.e., preferring *multi-modal* clarifications over *text-only* ones) with image quality (ρ=0.458), image clarity (ρ=0.626) and text clarity (ρ=-0.627). However, in Task 3, the user preference (i.e., preferring *multi-modal* clarifications over *visual-only*) had correlations only with the text clarity (ρ=0.505) and image clarity (ρ=-0.301). A closer look at the worker's feedback showed that the text and the image in more than 95% of clarification panes were relevant. This explained low to zero correlations between user preference and the relevance of the text and the image. We calculated the Tukey HSD test and observed the calculated correlations were significantly different from each other.

In the pairwise comparison between *multi-modal* and *visual-only* clarification panes, although the collected images for the clarification panes were the same, the workers preferred *multi-modal* clarification panes over the *visual-only* ones when the images were not clear. The text helped them understand the candidate answers to the clarification panes. The users preferred *visual-only* clarifications in more than 54% of cases when the text clarity was low and the image quality and clarity were high.

[8] Our results and codes are publicly available for reproducibility at https://github.com/Leila-Ta/MIMICS-MM.

[9] The Tukey HSD test is a post hoc test used when there are equal numbers of subjects in each group for which pairwise comparisons of the data are made [24].

Table 2
Motivations behind user preference (%).

Motivation	T vs. V		T vs. MM		V vs. MM	
	Prefer Text	Prefer Visual	Prefer Text	Prefer Multi-Modal	Prefer Visual	Prefer Multi-Modal
Easier to understand	25	31	7	61	6	67
Better decision	22	36	6	68	3	67
Faster decision	27	36	10	62	6	66
None of the above	8	12	4	9	9	5

However, the text and image were relevant in most cases.

In the post-task questionnaire, we investigated the users' motivation for their preferences. We asked users whether the preferred modality was easier to understand and helped them make better and faster decisions. Table 2 shows the user preferences in each pairwise modality. We see when users preferred *visual-only* clarification panes over *text-only* ones, 31% of users believed that the *visual-only* clarification panes were easier to understand The *visual-only* modality helped 36% of users make better and faster decisions. When comparing *multi-modal* clarification panes with *text-only* and *visual-only* clarification panes, between 60 to 70% of users believed that *multi-modal* clarification panes were easier to understand and helped them make better and faster decisions. Table 2 shows that there were small groups of users whose motivations behind their preferences were not listed in our questions.

User preference and impact of visual aspects. In the next step, we investigated the impact of visual aspects of the collected images on user preference over the clarification modality. We calculated the point-biserial correlation[10] [26] between the visual aspects of images and user preferences, the image quality and the image clarity. The average value of each aspect was calculated across all candidate answers for each clarification pane. Therefore, one value was obtained per visual aspect for every clarification pane. There was a low correlation between the image's visual aspects and user preference, including the image quality and clarity that the workers judged. To further explore the impact of visual aspects of images on user preference, we developed a feature-level attribution explanation to rate the image's visual characteristics based on their user preference. We utilised the Gini importance of the random forest with visual aspects as the input and target label user preference (i.e., 0 means Text preferred over Multi-Modal and one means Multi-Modal preferred over Text). The Gini importance is a metric that determines the relative significance of features in a random forest model. In this case, the visual aspects of the data were considered when calculating the Gini importance. By incorporating visual aspects into the Gini importance calculation, the model was likely able to capture and evaluate the relevance of visual features in the dataset. This can be particularly useful in scenarios where visual information plays a significant role or provides valuable insights for the given problem or task. We performed this analysis for Task 2, and the results indicate that *brightness*, *naturalness*, *RGB contrast*, *sharpness variation*, and *saturation variation*, among other studied aspects, accounted for more than 65% of the differences in user preferences. In particular, *brightness* and *naturalness* were the two most important visual features.

Automatic image generation for clarification panes. Finally, we investigated whether generating the corresponding images to the candidate answers could be automated. First, we compared the visual aspects (e.g., *brightness*, *colourfulness*, *naturalness*, ...) of the generated images with the collected ones. We observed that the generated images had relatively the same visual aspects as the collected ones. However, the Stable Diffusion model generated images with similar sharpness to the human-collected images.

Second, we compared computer-generated images with human-collected ones regarding image relevance, quality, and user preference. Table 3 shows that 87% of Stable Diffusion generated images

[10] The point-biserial correlation measures the relationship between a binary (i.e., user preference, image quality, and clarity) and a continuous variable (i.e., image aspects).

were relevant to the text. Even though only 20.7% of the generated images had a higher quality compared to human-collected ones, more than 57% of images had higher or equal qualities compared to collected ones. Only 12.7% of the generated images were preferred over the human-collected images. However, as seen from Table 3, 39.8% of the users either preferred the generated images or had no preferences over the generated and collected images (same preference). A slight improvement in the model performance was observed when we removed the irrelevant generated images from the collection (i.e., the percentage of generated images that had higher quality than the collected images rose from 20.7% to 21.2%, and the percentage of generated images that were preferred over collected images rose from 12.7% to 14.6%.). The annotators preferred the collected images over ~60% of computer-generated images. This observation was expected as the collected images were gathered through online searching to select the most suitable images. At the same time, a text-to-image model generated an image from only text. However, the Stable Diffusion model could generate relevant and high-quality images. As, in ~80% of cases, users preferred a *multi-modal* clarification pane over a *text-only* one; such a text-to-image model can ease and fasten the task of generating *multi-modal* clarification panes.

Table 3
Comparison of human-collected and computer-generated search clarification question images.

Collection Method	Relevance	Image Quality[1]	Image Preference[2]
Human-Collected	96%	42.7%	60.2%
Stable Diffusion model-Generated	87%	20.7%	12.7%

[1] 36.6% of users indicated the quality of the generated and collected images were the same. It is anticipated that by continuous improvements in the performance of text-to-image generation models, the image quality has significantly increased in the past two years.
[2] 27.1% of users indicated no preference over the generated and collected images.

5. Conclusions and Future Work

We aimed to understand the impact of clarification question modality on user preference. We introduced a novel *multi-modal* clarification dataset, *MIMCS-MM*. We created three modalities of *text-only*, *visual-only*, and *multi-modal* (a combination of both) for clarification panes and presented them to users through crowdsourcing.

The research shows that users generally preferred *multi-modal* clarification panes over *text-only* and *visual-only* ones. Users found it easier to understand the information presented in *multi-modal* panes, which helped them make better and faster decisions. This implies that integrating text and visual elements improves comprehension and decision-making for users, particularly given that the models for generating clarifications are not yet performing optimally. The study identified that when images were clear and of high quality, users favoured *multi-modal* panes. Therefore, ensuring that the visual content provided in clarification panes is of good quality and easily understandable is crucial. We also showed that when the images were unclear and of low quality, users preferred *text-only* clarification panes, even if the images were relevant. This suggests that when visual content is inadequate, relying solely on text can be more effective in conveying the necessary information.

We also explored the task of automatically generating corresponding images for *text-only* clarifications to make them *multi-modal* clarifications. The results indicated that text-to-image generation models, such as *Stable Diffusion*, can produce high-quality and relevant visual content. This indicates that automated generation techniques can produce *multi-modal* panes for search clarifications. Nonetheless, it is crucial to note that these methods have not yet achieved the ability to completely replicate human annotation when gathering relevant images for *text-only* clarification panes. Users still strongly prefer images collected by humans rather than those generated by models.

Our objective in this study was to gain insight into user preferences regarding different clarification modalities in a search scenario rather than examining the impact of clarification modality on search performance. As a result, we acknowledge that the participants in our study were not in a genuine search situation.

In our research, we recognise the potential impact of the dataset size. However, the statistically significant differences observed in our analysis form a reliable foundation for drawing valid conclusions. We have utilised robust statistical techniques to ensure the credibility of our findings, and it is unlikely that the observed effects are solely due to random chance.

The conducted study suggests several research paths for the future, including investigating the impact of clarification modality on search performance in real search situations, creating a more comprehensive dataset containing various aspects of queries to explore clarification modality further, developing advanced *multi-modal* language models to determine the most effective modality in different scenarios, investigating the impact of factors like user demographics, task complexity, and content characteristics, improving image generation techniques to produce more preferable images, and lastly, exploring alternative modalities beyond text and images, such as audio or interactive elements [27, 5]. We think future work should consider the development of more robust multi-modal clarification (e.g., images) using the latest advances in generative AI, large language models, and multi-modal foundation models, see [28, 29].

Acknowledgements

This research was supported in part by the Centre for Intelligent Information Retrieval, in part by the Office of Naval Research contract number N000142212688, and in part by NSF grant number 2143434. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] H. Zamani, J. R. Trippas, J. Dalton, F. Radlinski, Conversational information seeking, Foundations and Trends® in Information Retrieval 17 (2023) 244–456. URL: http://dx.doi.org/10.1561/1500000081. doi:10.1561/1500000081.

[2] H. Zamani, B. Mitra, E. Chen, G. Lueck, F. Diaz, P. N. Bennett, N. Craswell, S. T. Dumais, Analyzing and learning from user interactions for search clarification, in: Proceedings of SIGIR, 2020, p. 1181–1190.

[3] H. Zamani, G. Lueck, E. Chen, R. Quispe, F. Luu, N. Craswell, Mimics: A large-scale data collection for search clarification, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3189–3196.

[4] The alexa prize taskbot challenge, 2021. URL: https://www.amazon.science/alexa-prize/taskbot-challenge.

[5] Y. Deldjoo, J. Trippas, H. Zamani, Towards multi-modal conversational information seeking, in: Proceedings of SIGIR, 2021.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[7] P. Braslavski, D. Savenkov, E. Agichtein, A. Dubatovka, What do you mean exactly? analyzing clarification questions in cqa, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, 2017, pp. 345–348.

[8] L. Tavakoli, J. R. Trippas, H. Zamani, F. Scholer, M. Sanderson, Mimics-duo: Offline & online evaluation of search clarification, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3198–3208.

[9] M. Aliannejadi, H. Zamani, F. Crestani, W. B. Croft, Asking clarifying questions in open-domain information-seeking conversations, in: Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, 2019, pp. 475–484.

[10] J.-K. Kim, G. Wang, S. Lee, Y.-B. Kim, Deciding whether to ask clarifying questions in large-scale spoken language understanding, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 869–876.

[11] L. Tavakoli, J. R. Trippas, H. Zamani, F. Scholer, M. Sanderson, Online and offline evaluation in search clarification, ACM Trans. Inf. Syst. (2024). URL: https://doi.org/10.1145/3681786. doi:10.1145/3681786, just Accepted.

[12] J. R. Trippas, D. Spina, F. Scholer, Adapting generative information retrieval systems to users, tasks, and scenarios, in: R. W. White, C. Shah (Eds.), Information Access in the Era of Generative AI, Springer Nature Switzerland AG, Cham, Switzerland, 2024.

[13] S. Rao, H. Daumé III, Answer-based adversarial training for generating clarification questions, arXiv preprint arXiv:1904.02281 (2019).

[14] I. Sekulić, M. Aliannejadi, F. Crestani, User engagement prediction for clarification in search, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43, Springer, 2021, pp. 619–633.

[15] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, M. Li, Online video recommendation based on multimodal fusion and relevance feedback, in: Proceedings of the 6th ACM international conference on Image and video retrieval, 2007, pp. 73–80.

[16] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, Visual query suggestion, in: Proceedings of the 17th ACM international conference on Multimedia, 2009, pp. 15–24.

[17] M. Altinkaya, A. W. Smeulders, A dynamic, self supervised, large scale audiovisual dataset for stuttered speech, in: Proceedings of the 1st International Workshop on Multimodal Conversational AI, 2020, pp. 9–13.

[18] A. Srinivasan, V. Setlur, Snowy: Recommending utterances for conversational visual analysis, in: The 34th Annual ACM Symposium on User Interface Software and Technology, 2021, pp. 864–880.

[19] G. Pantazopoulos, J. Bruyere, M. Nikandrou, T. Boissier, S. Hemanthage, B. K. Sachish, V. Shah, C. Dondrup, O. Lemon, Vica: Combining visual, social, and task-oriented conversational ai in a healthcare setting, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 71–79.

[20] R. Ferreira, D. Silva, D. Tavares, F. Vicente, M. Bonito, G. Gonçalves, R. Margarido, P. Figueiredo, H. Rodrigues, D. Semedo, et al., Twiz: The multimodal conversational task wizard, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6997–6999.

[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).

[22] J. S. Hare, S. Samangooei, D. P. Dupplaw, Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images, in: Proceedings of the 19th ACM international conference on Multimedia, 2011, pp. 691–694.

[23] C. Trattner, D. Moesslang, D. Elsweiler, On the predictability of the popularity of online recipes, EPJ Data Science 7 (2018) 1–39.

[24] A. Stoll, Post hoc tests: Tukey honestly significant difference test, The SAGE encyclopedia of communication research methods (2017) 1306–1307.

[25] J. W. Tukey, Comparing individual means in the analysis of variance, Biometrics (1949) 99–114.

[26] R. F. Tate, Correlation between a discrete and a continuous variable. point-biserial correlation, The Annals of mathematical statistics 25 (1954) 603–607.

[27] J. R. Trippas, D. Spina, M. Sanderson, L. Cavedon, Towards understanding the impact of length in web search result summaries over a speech-only communication channel, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 991–994. URL: https://doi.org/10.1145/2766462.2767826. doi:10.1145/2766462.2767826.

[28] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, S. Milano, A review of modern recommender systems using generative models (gen-recsys), in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6448–6458.

[29] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, S. Milano, et al., Recommendation with generative models, arXiv (2024).