Improving Query and Assessment Quality in Text-Based Interactive Video Retrieval Evaluation

Werner Bailer
werner.bailer@joanneum.at
JOANNEUM RESEARCH
Austria

Rahel Arnold
rahel.arnold@unibas.ch
University of Basel
Switzerland

Vera Benz
vera.benz@unibas.ch
University of Basel
Switzerland

Davide Alessandro Coccomini
davidealessandro.coccomini@isti.cnr.it
CNR ISTI
Italy

Anastasios Gkagkas
gagastasos@iti.gr
CERTH ITI
Greece

Gylfi Þór Guðmundsson
gylfig@ru.is
Reykjavik University
Iceland

Silvan Heller
silvan.heller@unibas.ch
University of Basel
Switzerland

Björn Þór Jónsson
bjorn@ru.is
Reykjavik University
Iceland

Jakub Lokoc
jakub.lokoc@matfyz.cuni.cz
Charles University, Prague
Czech Republic

Nicola Messina
nicola.messina@isti.cnr.it
CNR ISTI
Italy

Nick Pantelidis
pantelidisnikos@iti.gr
CERTH ITI
Greece

Jiaxin Wu
jiaxin.wu@my.cityu.edu.hk
City University Hong Kong
Hong Kong

ABSTRACT

Different task interpretations are a highly undesired element in interactive video retrieval evaluations. When a participating team focuses partially on a wrong goal, the evaluation results might become partially misleading. In this paper, we propose a process for refining known-item and open-set type queries, and preparing the assessors that judge the correctness of submissions to open-set queries. Our findings from recent years reveal that a proper methodology can lead to objective query quality improvements and subjective participant satisfaction with query clarity.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Multimedia information systems**; • **General and reference →** *Evaluation.*

KEYWORDS

video retrieval, evaluation, benchmarking, quality assurance

ACM Reference Format:
Werner Bailer, Rahel Arnold, Vera Benz, Davide Alessandro Coccomini, Anastasios Gkagkas, Gylfi Þór Guðmundsson, Silvan Heller, Björn Þór Jónsson, Jakub Lokoc, Nicola Messina, Nick Pantelidis, and Jiaxin Wu. 2023. Improving Query and Assessment Quality in Text-Based Interactive Video Retrieval Evaluation. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3591106.3592281

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592281

1 INTRODUCTION

Recent advances in deep learning, among others through foundation models such as CLIP [6] and Florence [11], have brought significant improvement in image and video retrieval, in particular when the query involves translating between the text and image domains. However, fully automatic video retrieval is not suitable for all applications, such as when the information need of a user is fuzzy and cannot be easily phrased as a precise query, or when the domain is very specific and annotated data is too sparse to train powerful models. In these cases, interactive video retrieval systems, combining a number of search modes with browsing capabilities, are a powerful tool to quickly narrow down the content set to a relevant subset.

Evaluating the performance of these systems requires testing them as close as possible to real settings, with a human user in the loop. This can be best modelled in a live evaluation setting, where participating teams solve search tasks synchronously. For decades the dominant evaluation model for interactive video retrieval systems has been inspired by the Cranfield experiments in that they use a static evaluation paradigm [10]. Such benchmarks are used, for example, in TRECVID [1], VBS [8] and LSC [2]. Measuring precision and recall over fixed queries and data works well to some degree. The primary advantage of the static evaluation paradigm is the cost, as ground-truth annotations can be reused rather than using a prohibitively expensive battery of user evaluations. For any user-centred system evaluation approach, it is important to carefully construct these queries and evaluate task results. But in a live setting, in particular, there is little time to clarify any issues with the queries, and results should be available immediately after the end of the evaluation session.

The task types in interactive multimedia retrieval can be placed in a space spanned by different properties, such as the query presentation or the number of relevant results [5]. Two common types of video retrieval tasks are *Known-Item Search* (KIS) tasks and open-set tasks, for which we use the term *Ad-hoc Video Search* (AVS) tasks (following TRECVID and VBS). While KIS tasks have a single unique target segment in the entire dataset, which needs to be unambiguously specified by the query (e.g., visual content, a textual description), there is an undetermined number of relevant segments for AVS tasks, and the query is more general.

It has been observed that many relevant real-world video retrieval tasks, similar to AVS tasks, involve open result sets. Creating the ground-truth for the static evaluation of such queries is prohibitively expensive. Just imagine the time and effort needed to provide full annotations for a state-of-the-art video retrieval dataset, such as V3C [7] which contains thousands of hours of content. Thus, for live evaluations of interactive video retrieval systems, an alternative evaluation paradigm is needed. One such alternative is to assess the incoming results manually. This assessment needs to be distributed across a team of assessors (judges) in order to handle the possibly large number of submissions from the systems being evaluated. Due to the limitation in time and the number of assessors, having each submission assessed by more than one person is often not feasible. We note that the live assessment process has, for several years, received less attention than other aspects of the benchmark design.

In many practical video retrieval applications, queries are expressed in a text-based form. This makes bridging the semantic gap from textual description to a relevant segment of visual content an essential aspect of a video retrieval system. There are inherent limitations in finding a textual description of a (complex) visual scene that is understood in the same way by people with different backgrounds, e.g., in terms of culture, education, and native language. The interpretation of the query definition and the imagination of the scene that someone derives from it will strongly depend on that person's background. We cannot eliminate this issue, but we can reduce its impact on the evaluation of the retrieval systems as much as possible. Thus, two problems need to be addressed: (i) making the query formulation as clear and unambiguous as possible and (ii) ensuring that the assessors share the same understanding of the query (including possible edge cases) so that they will judge submissions consistently and fairly.

Both unclear formulations of queries and different interpretations of queries by participating teams and assessors have been observed in past benchmarks. There is a need for more accurate assessment and unambiguous query descriptions while keeping them as short, simple, and clear as possible, since most assessors and participants are non-native English speakers. This paper thus makes the following contributions in order to address these issues. (i) We describe a process for peer review and refinement of both textual known item (KIS-T) and open set (AVS) textual video queries, and for dry-running queries with the assessors. (ii) We provide an analysis of the changes made to the queries in this process for two editions of a benchmark. (iii) We evaluate the impact of the query refinement and assessor preparation process by comparing the agreement of teams with assessors' rulings. The contributions are made in two iterations of a benchmark. Participant feedback was collected via a survey after the evaluation conference.

The rest of this paper is structured as follows: In Section 2, we outline the process of refining queries and assessors' alignment, Section 3 shows a quantitative assessment of the alignment between assessors and participants and presents survey data showing how teams perceived the queries and assessments, and Section 4 closes with an outlook towards future interactive evaluation campaigns.

2 QUERY REFINEMENT PROCESS

The creation of the two types of queries starts with a single individual (called author henceforth) envisioning the query scenario and authoring the initial text description. A KIS-T query is constructed by picking a target clip from the dataset and then creating a three-part description that gradually reveals more details about the content of the scene. For AVS tasks, the author envisions a specific type of scene and checks that at least one instance is in the dataset (but there may be hundreds or even thousands relevant to this description). The description has a dual role as both guide for the participants and authority of correctness for the assessors.

In order to resolve ambiguities in the assessment, a briefing session for reviewing and revising queries has been introduced for the 2021 iteration, and complemented by a dry-run for open-set queries starting from the 2022 iteration.

2.1 Query Review and Discussion

An online meeting of all the assessors and the author is held to review and revise the queries. In the case of the KIS-T queries, the assessors first watch the short target video clip and then the text description is read out loud. Assessors give their comments on the wording of the description with the goal of improving the text.

The process for AVS queries also starts with the author reading the query description out loud. Then a discussion follows in order to clarify the author's true vision for the query and how the assessors understand the query description. The text is then jointly revised as needed. Revisions may include, for example, phrasing the query to make it more specific or wider, or to mention examples of what is considered in/out of scope (e.g., "balloons, not hot air balloons").

This meeting takes several hours as every single query is evaluated and revised (there are typically 20-25 queries in total). The changes to the text are documented in an online repository that is still available to the participants after the meeting.

2.2 Dry-Run for Open Set Queries

To further improve the quality of the AVS queries, a second revision meeting is held with all assessors and the author. This time the queries are not just discussed but actually put to the test in order to discover edge cases. A run using the revised AVS queries from the first meeting is organized using a fully functioning test-bed search system [9]. Multiple instances of the publicly available implementation[1] were hosted on a cloud machine. The participants in this meeting act as both searchers and assessors.

For each AVS query, the assessors first play the role of participants, trying to find video segments that not only satisfy the query

[1] https://github.com/siret-junior/somhunter

parameters but also determine the boundary of what should be accepted or not. Once sufficiently many results have been submitted (high tens) or enough time has passed, the author of the queries screen-shares the interface for judging submissions. Each submitted clip can then be discussed by the assessors with regard to the interpretation and the need for rephrasing. Exploring the queries and potential submissions in practice serves the dual purpose of getting a deeper understanding of the queries and synchronizing the evaluation process between the assessors in the real evaluation.

3 EVALUATION

We provide evaluation results by studying the changes that are made during discussion and dry-run, by analysing the agreement of assessors with participants before (2021) and after (2022) the proposed process was implemented, and measure the participants' perception of query clarity and assessment quality with an online survey in the 2022 and 2023, including questions about a comparison with the preceding year.

3.1 Impact on Queries

In order to analyse how the queries are changed through this process, we look both at the mean lengths of queries (Figure 1) and the number of changes (of selected types) per query (Figure 2), comparing the data for the 2022 and 2023 evaluations. As a general trend, we see that the lengths of queries increase in the discussion step and also slightly after the dry-run. This is a clear evidence that details and clarifications are added to the queries. It is interesting to note that the process seems to have improved overall in 2023: the initial queries (of both types) are longer than in the previous year, and the increase in length is smaller, meaning that fewer changes were deemed necessary. The final queries are even slightly shorter than in 2022.

As Figure 2 shows, the most common changes are adding or replacing nouns, adding or replacing adjectives, prepositions and similar words, and adding examples (positive or negative) to AVS queries. It can be observed that more changes are done to KIS-T than to AVS queries. However, there is no clear difference in terms of the types of changes between the query types or the sessions. Only examples seem to be added more frequently in the dry-run than in the discussion, which is understandable, as this is the time when specific corner cases are discovered. In line with the measured query lengths, the overall number of changes decreased in 2023 compared to 2022, indicating that the process ran smoother. However, there were still on average 1.38 word changes per query made in the process.

3.2 Agreement Between Teams and Assessors

Based on data analysis from two benchmark editions (2021 and 2022) published in [3, 4], we perform an analysis of the agreement between assessors and the participating teams. In 2021, only a discussion of queries was performed, while in 2022, the process described in this paper was introduced. Figure 3 plots the raw and weighted disagreement ratios. We analyse the agreement based on the number of teams submitting a particular video segment. As the teams interactively select these segments, a higher number of teams reflects a shared understanding of the relevance of the segment. Of course, some systems may have failed to retrieve some of these segments, even if the participants would consider them relevant to the query. The horizontal axis represents the number of teams t considering a particular video segment v_i as relevant (i.e., $\text{rel}_t(v_i) = 1$; 0 otherwise), and the vertical axis is disagreement ratio $r_t = \frac{\sum_i |\text{rel}_t(v_i) - \text{rel}_a(v_i)|}{\sum_i 1 - |\text{rel}_t(v_i) - \text{rel}_a(v_i)|}$, where $\text{rel}_a(v_i)$ is the assessors' decision. As t increases, the absolute number of video segments gets very low, and thus that metric becomes quite volatile. In addition, the absolute number of submissions differs between the years: the total number of submissions is 5,994 in 2021 but 10,161 in 2022. We thus propose to consider a weighted variant of this metric, normalised by the total number M of segments being assessed, i.e. $\hat{r}_t = \frac{t}{M} \frac{\sum_i |\text{rel}_t(v_i) - \text{rel}_a(v_i)|}{\sum_i 1 - |\text{rel}_t(v_i) - \text{rel}_a(v_i)|}$. Note that multiplying the disagreement factor by t rather than just normalising considers the fact that a segment is initially independently submitted by t teams. This metric is more comparable over the range of t.

In Figure 3 we can observe for both ratios that the number of disagreement cases drops quite significantly for segments found only by a single or two teams but does not drop or even slightly increases for video segments considered relevant by three or more teams. There are few cases in absolute numbers, but they still indicate that there are segments, where either the interpretation of the query differs, there are assessors' errors, or there are details in the video

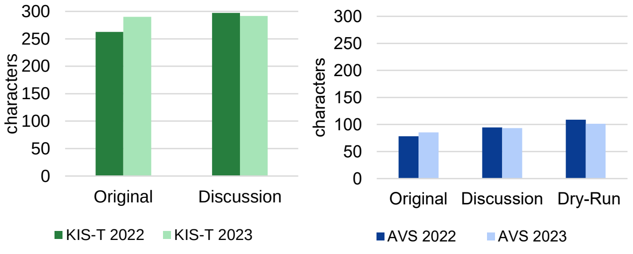

Figure 1: Mean query lengths in characters of text queries, comparing the originally proposed version, and the modified versions after discussion and dry-run (AVS only).

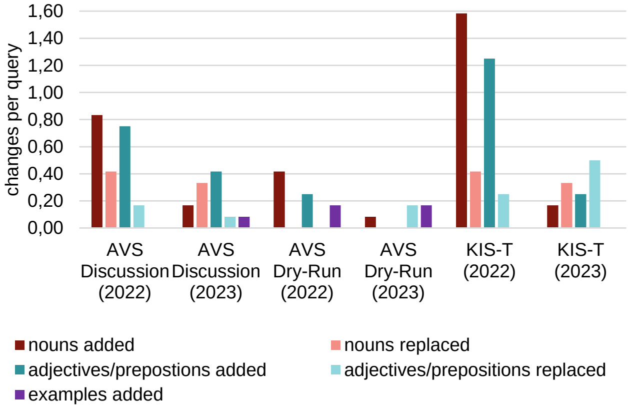

Figure 2: Selected types of modifications applied to the queries during discussion and dry-run (AVS only), expressed as number of changes per query.

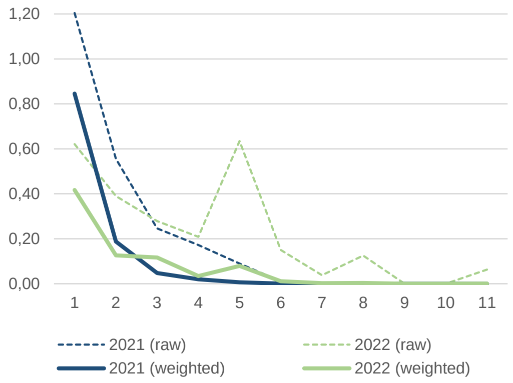

Figure 3: Raw and weighted disagreement ratios between assessment and t=1..11 teams.

segments, that are only apparent after careful inspection (as done by the assessors), and are missed by participants in the heat of the live evaluation. Nonetheless, the summed weighted disagreement ratio over $t = 1..11$ dropped by 0.318 from 2021 to 2022, which we attribute to the process described in this paper.

3.3 Participant Survey

In order to assess how participants perceived the impact of the preparation process on the quality of text queries and assessments, we performed a survey among the members of all participating teams. The survey was open the week after the 2022 and 2023 editions of the benchmark, respectively, using the same set of questions (provided as supplementary material). Participants filled in the survey anonymously, answering a set of multiple-choice questions and optionally providing additional comments. In both years, the survey was completed by about two-thirds of the participants ($n_{2022} = 20$, $n_{2023} = 18$). A majority of them were recurring participants ($n^r_{2022} = 14$, $n^r_{2023} = 13$), who were also asked to compare their experience with the previous year. Due to the different number of respondents in the two years, we present normalized rather than absolute numbers in the results.

When asked whether the textual queries are easy to understand (Figure 4), half of the participants responded with easy or very easy for KIS-T queries, and 90% for AVS queries. For both types of queries this share increases by 5% resp. 15% compared to 2022. It is worth noting that for AVS queries, no participant considered them difficult to understand, while this is the case for KIS-T queries, even if only for a few participants. This might be because all participants found at least one correct result for AVS tasks, which leads to a subjective satisfaction. When asked whether it is easy to decide whether a video segment matches the query, almost 80% found this easy or very easy, also an increase of about 15% compared to 2022 (although the share of very easy responses dropped slightly for unknown reasons).

Concerning perception of the assessment quality (Figure 5), two-thirds found the assessments very or mostly consistent, which is an increase of about 5% compared to 2022. Almost 80% found that their interpretation of the query was in line with that of the assessors (unchanged from 2022). It is worth noting that for both questions the share that gave the top rank (very consistent/all) nearly tripled from 2022 to 2023.

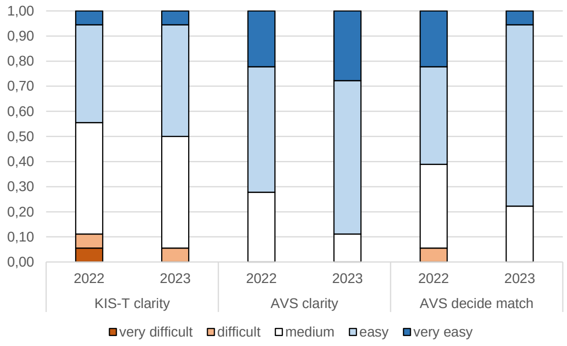

Figure 4: Assessment of clarity of KIS-T (left) and AVS (middle) queries: Participants rated the queries between very easy and very difficult to understand, and rated whether it was easy for AVS queries to decide if a video segment matches the query (right).

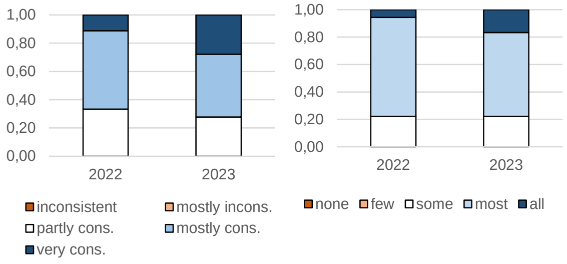

Figure 5: Assessment consistency (left, very consistent to inconsistent) and participants sharing the assessors' understanding of the queries (right, same understanding as assessors for none to all queries).

Finally, the repeating participants were asked to assess how the clarity of queries and the judgement quality has changed compared to the previous year (Figure 6). While the comparison 2021/2022 includes the introduction of the proposed process, the majority of participants still reports improvements for the comparison 2022/2023. For AVS query clarity and assessment quality, the share of participants answering better or much better slightly declined, still at almost 70%, but for both the share of much better slightly increased. While 60% already considered KIS-T queries better in 2022 than in the year before, this share increased to 85% in 2023, with 23% responding much better.

The participants of the survey could also provide additional comments. While a number of comments made in 2022 about the queries were positive (e.g., "*most of the AVS queries this year are easy to understand and clear*"), some participants criticized the structure ("*some sentences were quite nested and therefore difficult to understand*") as well as vocabulary, both in terms of language proficiency ("*some words were hard to comprehend as a non-native English speaker*") and semantics ("*meadow in my imagination was an Asian meadow*"). In 2023, there were much fewer comments, but

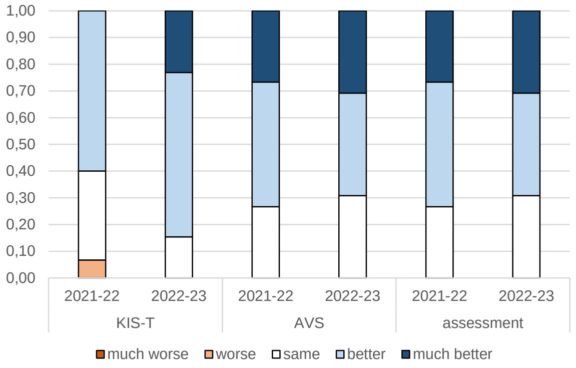

Figure 6: Comparative assessment of KIS-T and AVS query clarity and judging consistency between 2021/2022 and 2022/2023 by repeating participants.

there were some comments concerning assessment quality ("*I felt the judgements were much more consistent that in previous installments*"). In both years, some participants stated they found it hard to rate the assessments, as they did not get direct feedback, and cannot follow the large number of key frames on the scoreboard. One participant also proposed "*a feature where one could challenge a judgement verdict*".

4 CONCLUSION AND FUTURE WORK

In order to improve the quality of the evaluation of text-based queries in benchmarks for video retrieval systems, we have proposed a process for reviewing and revising the queries and preparing the assessors. We have analysed the changes made as result of this process, as well as the changes in the disagreement between participating teams and assessors. These data as well as the results of an online survey performed in two consecutive years show that the proposed process helps to improve the clarity of queries and the consistency of judgements.

However, there are still issues that can be considered in future work. A complete post-hoc reassessment of all submissions by multiple assessors per submission would provide detailed objective data to discover weaknesses in the process (this work is currently ongoing). For formulating the queries, the extent to which English language skills impact the performance could be studied, for example, by testing queries in a simpler version of the language (less precise, but maybe better understood by some), or by experimenting with machine/manually translated queries in the native language of participants. For open set queries, it would also be interesting to compare more specific and wider phrasings of the same query in order to understand which degree of simplification is possible.

ACKNOWLEDGMENTS

The research leading to these results has been funded partially by Cooperatio SCI-COMP, and the European Union's Horizon 2020 research and innovation programme, under grant agreement n° 951911 AI4Media (https://ai4media.eu), and under the Horizon Europe programme under grant agreement n° 101070250 XRECO (https://xreco.eu/).

REFERENCES

[1] George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. 2021. Evaluating multiple video understanding and retrieval tasks at TRECVID 2021. In *Proceedings of TRECVID Workshop*.

[2] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoć, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *International Conference on Multimedia Retrieval*. 685–687.

[3] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, et al. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal of Multimedia Information Retrieval* 11, 1 (2022), 1–18.

[4] Jakub Lokoč, Stelios Andreadis, Werner Bailer, Aaron Duane, Cathal Gurrin, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peška, Luca Rossetto, Loris Sauter, Konstantin Schall · Klaus Schoeffmann, Omar Shahbaz Khan, Florian Spiess, Lucia Vadicamo, and Stefanos Vrochidis. 2023. Interactive Video Retrieval in the Age of Effective Joint Embedding Deep Models: Lessons from the 11th VBS. under review.

[5] Jakub Lokoč, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Björn Þór Jónsson, Ladislav Peška, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, et al. 2022. A task category space for user-centric comparative multimedia search evaluations. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*. Springer, 193–204.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[7] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. 2019. V3C–a research video collection. In *International Conference on Multimedia Modeling*. Springer, 349–360.

[8] Klaus Schoeffmann. 2019. Video browser showdown 2012-2019: A review. In *International Conference on Content-Based Multimedia Indexing*. IEEE, 1–4.

[9] Patrik Veselỳ, František Mejzlík, and Jakub Lokoč. 2021. Somhunter V2 at video browser showdown 2021. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*. Springer, 461–466.

[10] Ellen M Voorhees. 2002. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*. Springer, 355–370.

[11] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).