IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 April 2024, accepted 20 May 2024, date of publication 27 May 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405638

TOPICAL REVIEW

Evaluating Performance and Trends in Interactive Video Retrieval: Insights From the 12th VBS Competition

LUCIA VADICAMO[1], RAHEL ARNOLD[2], WERNER BAILER[3], (Member, IEEE), FABIO CARRARA[1], CATHAL GURRIN[4], NICO HEZEL[5], XINGHAN LI[6], JAKUB LOKOC[7], SEBASTIAN LUBOS[8], (Graduate Student Member, IEEE), ZHIXIN MA[9], NICOLA MESSINA[1], THAO-NHU NGUYEN[10], LADISLAV PESKA[7], LUCA ROSSETTO[11], LORIS SAUTER[2], KLAUS SCHÖFFMANN[12], (Member, IEEE), FLORIAN SPIESS[2], MINH-TRIET TRAN[13], AND STEFANOS VROCHIDIS[14], (Member, IEEE)

[1]Institute of Information Science and Technologies, National Research Council (CNR), 56124 Pisa, Italy
[2]Department of Mathematics and Computer Science, University of Basel, 4051 Basel, Switzerland
[3]Joanneum Research, 8010 Graz, Austria
[4]ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
[5]HTW Berlin–University of Applied Sciences, 10318 Berlin, Germany
[6]School of Computer Science, Wuhan University, Wuhan 430072, China
[7]Faculty of Mathematics and Physics (FMP), Charles University, 118 00 Prague, Czech Republic
[8]Institute of Software Technology, Graz University of Technology, 8010 Graz, Austria
[9]School of Computing and Information Systems, Singapore Management University, Singapore 188065
[10]School of Computing, Dublin City University, Dublin 9, Ireland
[11]Department of Informatics, University of Zurich, 8050 Zürich, Switzerland
[12]Institute of Information Technology, Universität Klagenfurt, 9020 Klagenfurt, Austria
[13]Software Engineering Laboratory, Faculty of Information Technology, University of Science–VNUHCM, Ho Chi Minh City, Vietnam
[14]Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece

Corresponding author: Lucia Vadicamo (lucia.vadicamo@isti.cnr.it)

This work was supported in part by the European Commission through the projects "AI4Media–A European Excellence Centre for Media, Society and Democracy" under Grant EC, H2020 951911, "SUN–Social and hUman ceNtered XR" under Grant EC, Horizon Europe 101092612, and "XRECO–XR mEdia eCOsystem" under Grant EC, Horizon Europe 101070250; in part by the Swiss National Science Foundation projects "Participatory Knowledge Practices in Analog and Digital Image Archives" under Contract 193788, and "MediaGraph" under Contract 202125; in part by the Austrian Research Promotion Agency (FFG) under Project 886205; in part by the Vingroup Innovation Foundation (VINIF) under Project VINIF.2019.DA19; in part by Czech Science Foundation (GAČR) under Project 22-21696S; in part by the FWF Austrian Science Fund under Grant P 32010-N38; in part by the National Natural Science Foundation of China under Grant U1903214, Grant 62372339, and Grant 61876135; and in part by the Science Foundation Ireland under Grant numbers SFI/13/RC/2106_P2, SFI/12/RC/2289_P2 and 18/CRT/6223.

ABSTRACT This paper conducts a thorough examination of the 12th Video Browser Showdown (VBS) competition, a well-established international benchmarking campaign for interactive video search systems. The annual VBS competition has witnessed a steep rise in the popularity of multimodal embedding-based approaches in interactive video retrieval. Most of the thirteen systems participating in VBS 2023 utilized a CLIP-based cross-modal search model, allowing the specification of free-form text queries to search visual content. This shared emphasis on joint embedding models contributed to balanced performance across various teams. However, the distinguishing factors of the top-performing teams included the adept combination of multiple models and search modes, along with the capabilities of interactive interfaces to facilitate and refine the search process. Our work provides an overview of the state-of-the-art approaches employed by the participating systems and conducts a thorough analysis of their search logs, which record user interactions and results of their queries for each task. Our comprehensive examination of the VBS competition offers assessments of the effectiveness of the retrieval models, browsing efficiency, and user query patterns. Additionally, it provides valuable insights into the evolving landscape of interactive video retrieval and its future challenges.

The associate editor coordinating the review of this manuscript and approving it for publication was Tomas F. Pena.

INDEX TERMS Content-based retrieval, interactive evaluation campaign, interactive video retrieval, performance evaluation, video browsing, video content analysis.

I. INTRODUCTION

"A picture is worth a thousand words", goes the age-old adage, yet the challenge is that we do not always have access to the perfect image to convey our message. The image itself may be the primary target of our information need, requiring alternative means of expression. For example, consider a journalist searching for a specific video within a vast, unannotated multimedia collection based on a fleeting memory. If even a single frame of the desired video were readily available, finding the complete video would become trivial in the realm of computer vision and video retrieval. However, in practice, the journalist must rely on alternative means to describe the content of the target video. While words serve as the most immediate and utilized tool for conveying descriptions, visual cues such as similar images or sketches can also be used fruitfully.

In recent years, significant efforts have been made to develop high-performance video retrieval systems, allowing users to employ various search capabilities, including text, visual, or multimodal queries. These systems may also actively engage users in the search process, allowing them to refine queries, explore results, provide feedback, and iteratively navigate the video content to fulfill their information needs. However, assessing and comparing these systems poses a significant challenge due to their interactive nature and diverse supported query and search modes, making it impractical to conduct a static comparison against a conventional benchmark dataset. To address this, live benchmarking campaigns, such as the Video Browser Showdown (VBS) [56], [74] and the Lifelog Search Challenge (LCS) [51], [109], have emerged as crucial initiatives.

This paper provides an in-depth evaluation of the 2023 iteration of VBS, an international video content search competition held annually since 2012 at the International Conference on Multimedia Modeling (MMM). It has become a well-established benchmark offering comparative insights into state-of-the-art interactive video search systems. During VBS, participants face two main tasks: Known-Item Search (KIS) and Ad-hoc Video Search (AVS), both of which are to be completed within a predefined time limit. KIS tasks require participants to locate a specific video clip within a dataset, with the instance either visually presented or described textually by a moderator. In AVS tasks, participants must find as many video clips as possible that match a general textual description. Scoring considers factors such as search time, false submissions, and the diversity of instances found in AVS tasks.

Since VBS search tasks [72] require automatic analysis of general video content and the "out-of-the-box" effectiveness of video retrieval models, the competition's impact extends beyond evaluation, actively shaping the evolution of interactive video search systems. Winning approaches often set the tone for the coming years, guiding researchers and developers towards promising directions. For instance, based on the winning approaches of the last few years, mirroring trends in the broader computer vision domain, we observed an indisputable shift not only from traditional handcrafted similarity search models to modern deep learning approaches [50] but also a move towards the prevalent use of joint embedding models [67], [86]. These models have played a central role in current research on interactive video retrieval due to their capability to integrate information from multiple modalities, along with enhanced semantic understanding. Moreover, they exhibit a notable capacity to generalize across domains and tasks. Training these models on extensive web-scale datasets enhances their ability to efficiently search for a diverse range of concepts, as well as whole phrases linking concepts with additional properties (e.g., "blue bird on a branch", "white shirt and blue jeans"). Consequently, it is not surprising that in VBS 2023, most of the teams used a CLIP model [60], [86] or other multimodal embeddings [66], [77], [81] in their interactive video search systems. Teams integrating recently introduced versions of the CLIP model trained on LAION datasets [60], [101] demonstrated impressive performance, while other teams with the original CLIP [86] remain competitive. However, our analysis emphasizes that interactive interfaces designed on top of the corresponding multimodal ranking models are crucial. Systems that used the same latest CLIP model did not perform consistently. Factors differentiating system effectiveness include the browsing interface, the ability to combine different models, and the option to reorder results based on temporal searches and visual similarity.

The main contributions of this paper can be summarized as follows:

- Providing a valuable comparative analysis of systems that participated in the 12th VBS competition, outlining the state-of-the-art approaches adopted and illustrating the latest trends in interactive video retrieval.
- Offering a comprehensive overview of the competition settings and outcomes, including overall scores, the number of correct and incorrect submissions, and submission times for each team and task.
- Delving into an in-depth analysis of teams' performance during KIS tasks, offering assessments of retrieval model effectiveness, browsing efficiency, and user query patterns.
- Exploring the outcomes of AVS tasks, including timeline statistics, success rates, task difficulty analysis, and agreement with judges' assessments.
- Providing a critical analysis of current challenges and suggesting pathways for future improvements in upcoming VBS evaluations.

The remainder of this paper is structured as follows: Section II summarizes the VBS 2023 settings and tasks;

FIGURE 1. VBS2023 featured a hybrid format. In-person teams gathered in a room with a large screen where tasks were projected. Meanwhile, online teams accessed the main DRES overview through their web browsers, where tasks and scores were displayed.

Section III outlines approaches utilized by the participating systems; Section IV provides a comprehensive analysis of system results and queries during the VBS tasks; Section V delves into current challenges in interactive VBS evaluation and provides recommendations for future VBS editions; Section VI discuss our findings, future challenges, and research directions in interactive video search; Section VII draws the conclusions.

II. COMPETITION SETUP

VBS is a live competition to evaluate interactive search tools. Over the last few years, it has employed a subset of the Vimeo Creative Commons Collection (V3C [93]), comprising 17,235 video files totaling 2,300 hours of video content. In 2023, the competition introduced the use of the Marine Video Kit (MVK [111]), a smaller but very challenging dataset. The MVK consists of 1,374 videos, amounting to approximately 12 hours, showcasing underwater scenes.[1]

VBS 2023 was a hybrid event in which both in-person and online teams participated (Fig. 1). Each team consisted of a maximum of two operators authorized to operate the retrieval system individually. The competition was controlled by the DRES evaluation server [90], which controlled task presentation and collection and evaluation of submissions.

Similarly to previous editions, VBS 2023 included KIS and AVS tasks. For KIS tasks, there exists a single unique correct segment in the collection. The query is either presented as the target video segment (referred to as KIS-V, representing visual KIS) or as a textual description of the contents of this segment (referred to as KIS-T, representing textual KIS), which usually extended during the working time. In contrast, AVS queries are broader textual queries with an undetermined number of correct items. The ground truth is thus not a priori-defined but established during the competition using live judging. Some KIS-V queries used the MVK dataset (referred to as KIS-V-M), while all other tasks used the V3C dataset. Each task has a limited working time (7 minutes for KIS-T, 5 minutes for others), with penalties for incorrect submissions. Submissions are assessed against the ground truth for KIS tasks, whereas for AVS, they are assessed by live judges. The scoring for the KIS tasks, detailed in [56], involves rewarding speed in finding the correct item while penalizing wrong submissions. For AVS, a new scoring formula was applied to foster a diversity of submissions. Teams receive scores for the first correct submission of each video, and a penalty is added for wrong submissions to prevent the submission of unverified shots. The score f_t of a team t is determined as in [70]:

$$f_t = 1000 \cdot \max\left(\frac{1}{|C|}\sum_{v}^{\mathcal{V}_t}(c_v - i_v p), 0\right), \text{ where}$$

$$\begin{aligned} C &:= \text{set of correct videos across all teams' submissions} \\ \mathcal{V}_t &:= \text{set of videos with a submission for team t} \\ c_v &:= 1 \text{ if there is a correct submission for } v, 0 \text{ else} \\ i_v &:= \text{number of incorrect submissions before the first} \\ &\quad \text{correct submission for video } v \\ p &:= \text{submission penalty constant (set to 0.2)} \end{aligned} \tag{1}$$

This score function considers diversity, as submitting multiple correct items from the same video does not increase a team's score. Furthermore, submissions associated with videos not commonly discovered by other teams can significantly impact the evaluation, given that each team's score is divided by the total number of correct videos from all teams.

The challenge in the query formulation process is not only to ensure that the content exists in the collection and – in the case of KIS queries – there is a unique target that can be unambiguously identified but also to ensure that queries are clearly phrased and also understandable by non-native speakers. This includes, for example, choosing between terms that would describe an object more precisely vs terms that are more commonly used and broadly understood. An additional challenge for AVS queries is to ensure a common understanding of the judges of how to interpret the query and how to treat border cases that arise. This is important to ensure consistent judgment of all submissions. In 2022, a process for reviewing queries with the team of judges and performing a dry run for AVS queries has been introduced and repeated for VBS 2023. Details on the process and its evaluation can be found in [36]. A survey among participants confirms that the goals of improving query quality and judgment consistency are reached with this process.

III. OVERVIEW OF STATE-OF-THE-ART APPROACHES USED BY PARTICIPATING SYSTEMS

This section provides a concise overview of the advanced techniques employed by participating systems in the VBS 2023 (Table 1). It highlights the latest advancements in video

[1]Note that a snapshot of the dataset from 2022, modified to the needs of VBS, has been used.

TABLE 1. List of participating systems and selected approaches used by them. The systems are ranked by their overall score in VBS 2023, with the "Solved KIS" column indicating the number of tasks completed out of the 19 KIS tasks issued during the competition. In the "Shot detection" columns, the symbol "∗" denotes the utilization of predefined shots from the V3C dataset [93]; when a time value is present, it indicates the application of uniform sampling with the specified time interval; otherwise, when available, a reference to the method used is provided. The symbol ✓ indicates that a method is used. A light gray color indicates that the feature is present but was not (or just rarely) used. In the "Joint Embedding" column, the symbols ✓, $✓^2$, and $✓^3$ correspond to the usage of one, two, and three multimodal embedding models, respectively. The ASR data for V3C was provided by [92].

System info					Shot detection		Search								Browsing					
	Country	Overall score	Solved KIS	No. Users	V3C dataset	MVK dataset	Joint Embedding	Concepts	ASR	OCR	Query-By-Example	Temporal Query	Relevance Feedback	Other	Top-k video filter	Temporal context	Video Summary	Video Player	Video Preview	2D map embedding
vibro [99]	DE	3,992	18	2	[58]	[58]	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓
VISIONE [31]	IT	3,625	17	2	*	1s	$✓^3$	✓			✓	✓		✓	✓	✓	✓	✓	✓	
VIREO [79]	SG	3,258	16	2	*	3s	$✓^2$	✓	✓	✓	✓	✓				✓	✓	✓	✓	
vitrivr-VR [107]	CH	3,200	16	2	*	1s	$✓^2$					✓								
CVHunter [71]	CZ	3,027	13	2	custom	custom	✓				✓	✓	✓		✓	✓	✓		✓	
vitrivr [96]	CH	2,986	14	2	*	1s	$✓^3$	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓	
Verge [84]	GR	2,803	13	2	*	1s	$✓^3$	✓			✓	✓		✓		✓	✓	✓	✓	
QIVISE [103]	CN	2,314	14	1	*	2s	✓				✓		✓		✓	✓	✓			
VideoCLIP [82]	IE	1,858	9	2	*	1s	✓	✓		✓	✓			✓	✓	✓	✓	✓	✓	
v-FIRST [59]	VN	1,773	9	1	*	1s	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓	
diveXplore [100]	AT	1,647	9	1	[104]	10s	✓	✓		✓	✓			✓	✓		✓	✓		
4MR [34]	CH	1,626	10	2	*	1s	✓				✓				✓			✓	✓	
Perfect Match [76]	AT	34	0	1	*	2s		✓		✓					✓					

search and retrieval, showcasing the progress made by the research community in improving video exploration and analysis. It covers important subsections such as Joint text-visual embedding methods, Concept Search, Query by Example, Temporal Querying, Relevance Feedback, and Browsing. The repositories and relevant scientific articles have been referenced for the open-source approaches utilized. It's important to note that among the systems, only VISIONE, vitrivr, and vitrivr-VR are open source [4], [8], [23], [25], [26].[2] The other systems, although not entirely open-source, use many models and approaches published in open-source repositories, which are referred to in the following sections.

A. JOINT TEXT-VISUAL EMBEDDING METHODS.

The VBS systems have greatly evolved in recent years, offering innovative approaches to explore and retrieve information from large video collections efficiently. Almost all these systems exploit joint text-visual embeddings to enhance the search experience and provide more accurate results. We can broadly categorize these systems into groups based on the number of multimodal embedding models they employ: those utilizing a single model and those using multiple models. In Table 1, the symbols ✓, $✓^2$, $✓^3$ represent the usage of one, two, and three multimodal embedding models, respectively. The specific models used by each system are summarized in Table 2.

[2]vitrivr, and vitrivr-VR comprise three components: the user interface [25], [26], the Cineast retrieval and feature extraction engine [4], and the Cottontail database [8].

TABLE 2. Employed joint text-visual embedding models. The solid line separates CLIP-based models from other multimodal embedding models.

Model	System
OpenCLIP ViT-L/14@336 trained with LAION-2B [15], [60]	VideoCLIP [82] v-FIRST [59]
OpenCLIP ViT-L/14 trained with LAION-2B [14], [60]	vibro [99] VISIONE [31]
OpenCLIP ViT-L/14 trained with LAION-400m [60]	diveXplore [100]
OpenCLIP ViT-B/32 trained with LAION-2B [12], [60]	4MR [34]
OpenCLIP ViT-B/32 xlm roberta base model trained with LAION-5B [13], [60]	vitrivr [96] vitrivr-VR [107]
CLIP [5], [86]	CVHunter [71] vitrivr [96] vitrivr-VR [107]
CLIP2Video [6], [45]	VISIONE [31]
BLIP [3], [66]	QIVISE [103]
CLIP4Clip [7], [77]	VIREO [79]
Custom cross-modal network [20], [46] combining multiple textual and visual features and employing OpenCLIP ViT-B/32 [60], [86], ResNet-152 [53], and ResNeXt-101 [80]	Verge [84]
ITV [116]	VIREO [79]
ALADIN [2], [81]	VISIONE [31]
custom model [24], [105]	vitrivr [96] vitrivr-VR [107]

Several notable implementations stand out in the category of VBS systems with a single model. For instance, vibro [99] employs the OpenCLIP ViT-L/14 [14], [60] trained on LAION-2B [101] to produce joint text-visual embeddings. VideoCLIP [82] and v-FIRST [110] uses the visual transformer CLIP ViT-L@336 [15], [60], [86] trained on the LAION-2B dataset. In VideoCLIP, the integration

of Milvus [113] vector database facilitates seamless matching between embeddings. v-FIRST [59] presents a revised version of their previous interactive video retrieval system [110], which supports querying by textual descriptions and visual examples. The joint text-visual feature space is the basis for many of v-FIRST's functionalities, such as optimized vector search, fast neighbor search, and compression of similar video segments. diveXplore [100] leverages the OpenCLIP ViT-L/14 model trained on LAION-400m [60] to extract visual embeddings from keyframes. The embeddings are indexed in a FAISS [63] index that is used for all free-text queries by a Python server running in the backend. This server extracts embeddings from a text query, compares them with an L2 distance to the visual embeddings of the keyframes, and returns the ranked results via a WebSocket connection to the frontend. 4MR [34] also uses a CLIP model, the ViT-B/32 [12], [60], [86] pre-trained on LAION-2B. A Python server in the backend transforms the input to a vector, which is afterward used for similarity search. QIVISE [103] employs the BLIP [3], [66] model (ViT-B and CapFilt-L, the one trained on 129M images), an advancement built on the foundations of the CLIP [5], [86] model. In QIVISE, the BLIP model is used to extract feature vectors from both textual queries and images, then computes the cosine similarity between these vectors. CVHunter used a CLIP [86] model as well. However, the original version (i.e., not trained with LAION data) was used.

On the other hand, VBS systems that utilize multiple joint embeddings employ a range of sophisticated techniques to enhance the search process. For instance, VISIONE incorporates three models: CLIP ViT L/14 [14], [86] trained on the LAION-2B dataset, CLIP2Video [6], [45], and ALADIN [2], [81], its own cross-modal model. ALADIN generates high-quality scores by aligning images and texts using a pre-trained vision language transformer and then trains a shared embedding space using a cross-modal alignment head. The VISIONE system effectively combines the results of these three models using a late fusion algorithm. The CLIP and CLIP2Video features are indexed and searched using FAISS library [63], while ALADIN features are transformed into a textual format (Surrogate Text Representation [32]) to be indexed and searched using Apache Lucene.[3] Similarly, other systems like vitrivr and vitrivr-VR [96] rely on custom visual-text co-embedding [24] techniques (similar to W2VV++ [67]), along with CLIP and OpenCLIP [60], [101] models (xlm-roberta-base-ViT-B-32 using pre-trained laion5b_s13b_b90k weights [13]), providing multilingual query support and enabling the search for videos using natural language prompts. One of the benefits of OpenCLIP is its multi-language model, which empowers users to formulate queries in a lot of different languages, such as, but not limited to, English and German. Verge [84], on the other hand, utilizes three distinct trained networks, namely ResNet-152 [53], ResNeXt-101 [80], and the CLIP model ViT-B/32 [60], [86], to perform text-to-video matching [20], [46]. Converting the intricate textual query and videos into a shared latent space allows direct comparison. Subsequently, an attention-based dual encoding network is utilized. Four extensive video caption datasets (MSR-VTT [118], TGIF [68], ActivityNet [39], and Vatex [115]) were used to train the model. VIREO expands the embedding bank with CLIP4Clip feature [7], [77] based on the previous system [78] which relies mainly on the ITV feature [116]. The CLIP4Clip feature is fine-tuned on the MSR-VTT [118] dataset. In addition, the late fusion of different features is also used to diversify the search results.

B. CONCEPT SEARCH

Concept search enhances video retrieval by allowing users to search for videos based on specific concepts or semantic information. The participating systems in VBS 2023 employed various techniques, such as keyword decomposition, concept probability estimation, and pixel-wise concept annotation.

Over the years, vibro has been at the forefront of incorporating text-based methods for video search, including OCR, ASR, and automatic annotations of frames. However, interestingly, none of these text-based methods were utilized in the VBS 2023 challenge.

In contrast, the VISIONE system, similar to its previous version [30], focused on object detection using three deep convolutional neural network models: VarifocalNet [11], [119], Mask R-CNN [11], [52], and Faster R-CNN [10], [48]. These models were trained on different datasets, namely COCO, LVIS, and OpenImages v4. To ensure consistency and organization of class labels, the VISIONE system implemented a hierarchical structure based on WordNet.[4]

VIREO adopted concept search as a complementary approach to embedding search. Through keyword decomposition and concept probability estimation [116], VIREO provided a ranked list of video shots associated with each concept.

vitrivr employed pixel-based color [47] and concept search [91] methodologies, like in previous iterations [55]. It leveraged DeepLab pixel-wise concept annotation [40] and implemented post-processing techniques, such as resolution reduction and label transformation, to facilitate efficient and effective spatially localized concept search. Additionally, vitrivr introduced the concept of Query-by-Semantic-Sketch, allowing users to search using a concept brush.

diveXplore [100] offered search capabilities for visual concepts using EfficientNet [108], which was trained on datasets such as Places365 [120], ImageNet-1K [65], and GPR1200 [97]. However, the utilization of concept-based search was limited in the challenge due to the superior performance of CLIP, which overshadowed other methods.

Verge [84] continued to build upon its previous version [33] by employing a 3D-CNN architecture for

[3]https://lucene.apache.org/

[4]Available at https://zenodo.org/records/7194300

spatio-temporal human activity recognition [1]. The system followed a three-step pipeline [49] involving object detection, object tracking, and activity recognition. This allowed Verge to identify and recognize human-related activities in videos effectively. Moreover, the system exploits Yolo v4 [27] for human and face detection, WideResNet and ResNet50 pretrained models [18] for Places365 concept detection, and EfficientNetV2-L pretrained model [9], [21] for ImageNet concept detection, Pretrained Sentence-BERT model (stsb-mpnet-base-v2) [16] for concept label similarity inference.

v-FIRST indexes different concepts and allows users to apply Boolean retrieval for added flexibility [59]. This combination of concept search and joint embedding is implemented in their unified database.

VideoCLIP inherited features from its previous version [83] and incorporated K-means clustering for dominant color determination at a pixel level. It also employed the Yolov5 model [62] for extracting visual concepts from videos. These features enhanced the search capabilities of VideoCLIP, enabling users to retrieve videos based on color and visual concepts.

Finally, Perfect Match utilized various concept detection models [62], [65], [86] and precomputed results for efficient frame-level searching. By leveraging different classification datasets [38], [69], [94], [117], such as ImageNet, MSCOCO, Food-101, and SUN397.

C. QUERY BY EXAMPLE

Many VBS systems support query-by-example, allowing users to use an image or video frame as a query to discover visually or semantically similar content. Similar to cross-modal search, where a text prompt is used to search for a video, we observed that the prevailing approaches for visual similarity search rely on features extracted from the visual transformer of multimodal models like CLIP.

For example, vibro [99] employs a CLIP ViT-L [15], [86] network that has been pre-trained on the LAION-2B dataset and fine-tuned in publicly available image datasets [98]. This enables the system to extract feature vectors useful for content-based image retrieval [97]. On the other hand, VISIONE provides support for visual and semantic similarity searches. It utilizes GEM features [88] for visual similarity search and incorporates CLIP2Video [6], [45] and ALADIN [2], [81] for searching semantically similar video clips. CVHunter and VideoCLIP utilize the same CLIP features for image similarity search as they do for text-to-image search. VIREO measures the similarity between a shot query and all shot candidates using the fine-tuned CLIP4Clip [77] feature. diveXplore [100] uses the CLIP ViT-L/14 model to extract image embeddings from an example image, which are then sent to the Python server in the backend to query the FAISS [63] L2-index for keyframes with similar embeddings. Verge [84] incorporates a visual similarity search module that facilitates the retrieval of visually similar content based on a query image. This module utilizes feature vectors generated from a fine-tuned GoogleNet architecture [85] and leverages an efficient IVFADC indexing structure [61]. v-FIRST employs optimized nearest neighbor algorithms in the embedding subspace [59] to identify targets similar to the example image or text. Furthermore, in v-FIRST, an image generator based on MidJourney and Stable Diffusion is integrated to synthesize images from a text prompt as an additional query methodology. vitrivr and vitrivr-VR use simple color and edge features for query-by-example. At any point, while watching a video, the current frame can be used as a source image for query-by-example using these features. 4MR [34] employs the CLIP model ViT-B/32 [12], [86] for query-by-example. In an offline phase, all keyframes were extracted beforehand. These CLIP feature vectors are used to retrieve objects similar to a given example.

D. TEMPORAL QUERYING

Temporal queries are crucial to enhancing VBS systems' search capabilities, with many incorporating this functionality to facilitate users in searching for specific patterns or relationships within video clips.

For instance, VISIONE enables temporal queries by describing two scenes from the same video clip. It utilizes a temporal quantization approach, dividing video time into intervals and independently processing the results of both queries to retain representative results for each time interval and query. Result pairs from the same video with a temporal distance smaller than a certain threshold are displayed as the temporal search results.

Similarly, CVHunter and vibro utilize a temporal query fusion technique, computing two arrays of scores for different temporal query parts and then fusing them to generate final results. In vibro, keyframes are extracted at a rate of 2 frames per second, with temporal queries considering only temporally close keyframes. CVHunter uses a similar frame extraction approach as VIRET [73].

VIREO supports temporal queries consisting of two successive and independent queries. A sliding window approach is used to aggregate the scores of the two queries and index the shots between shot pairs that match the two queries.

In Verge [84], temporal queries are restricted to concepts, enabling users to search for two concepts appearing consecutively within the same video. The system generates separate shot probability lists for each concept, calculates the intersection of concepts, and re-ranks shots using an objective function.

Both vitrivr-VR and vitrivr offer temporal querying capabilities, allowing users to search for specific patterns or relationships in consecutive video segments. These systems enable the combination of multiple non-temporal queries into a single temporal query.

v-FIRST facilitates finding two sequential images in a video by summing the embeddings of each image to create

a new representation and then searching within the collection of embeddings for all possible pairs.

E. RELEVANCE FEEDBACK

In the realm of user relevance feedback [44], [54], [64], [95], several VBS systems have exploited innovative approaches to enhance the retrieval process based on user interactions.

CVHunter incorporates the Bayesian relevance feedback model [43], allowing users to provide feedback on the relevance of retrieved video clips. In this framework, the system maintains probability for each image in the database, estimating its relevance to the user. In each iteration, example images are provided in addition to a list of selected implicit negative examples. CVHunter improved this system by providing a temporal version of the Bayesian feedback. This model enables feedback for the single and temporal variant [75], empowering users to refine their queries and obtain more accurate search results.

QIVISE introduces a novel quantum-inspired interaction paradigm for modeling user interactions. Building upon recent studies highlighting the potential of quantum theory's mathematical framework for information retrieval [112], QIVISE integrates state-of-the-art quantum-inspired re-ranking paradigm [114] along with feedback processing methodologies [41]. After the initial retrieval phase, users can select video clips that are highly consistent or inconsistent with their demands. Within a quantum state space, these selected clips are then used to estimate the user's actual demands, utilizing the space spanned by the chosen clip vector and its complement subspace. For the final re-ranking score, unlike the Rocchio Algorithm [42], QIVISE utilizes the relevance probability from the previous retrieval round instead of treating the relevance probabilities of all selected video clips as equal. For a detailed explanation of user demand estimation and re-ranking score calculation, refer to the QIVISE paper [103].

In vibro, users have a dedicated application window for AVS tasks. Any video frame selected and sent to the evaluation server is considered positive feedback to the system [99]. The integrated feedback loop computes the minimum distance of all the previously selected frames and the remaining frames in the dataset. The results list will be displayed in the same AVS windows, and the user can repeat the process as often as desired.

v-FIRST implements an optional query reweighting using the top retrieved images [59]. This approach helps mitigate irrelevant factors and emphasizes concepts that are crucial to the retrieval process.

F. OTHER

In the context of incorporating additional features and modules into VBS systems, several approaches have been adopted to enhance the search experience and provide more comprehensive results.

VISIONE utilizes the Whisper model [19], [87] to integrate a speech-to-text feature into the system. This feature allows users to dictate their queries rather than type them, making the process of issuing textual queries more convenient. The spoken text is automatically translated into English and used as a query for the cross-modal search modules.

CVHunter enhances the result set by augmenting it with labels assigned through zero-shot CLIP classification. A pre-selected set of class labels is used to classify the images in the result set. These labels are displayed below each image in the main search panel, providing users with additional information about the content and allowing them to learn words associated with various images, as assigned by CLIP.

diveXplore [100] incorporates text detection and recognition features using the CRAFT model [17], [35]. This model is utilized to detect regions with text in keyframes and subsequently recognize the text in those regions. Additionally, the YOLOv5 [28], [62] model is employed to detect COCO objects in the keyframes, enhancing the system's understanding of the visual content. Keyframes, extracted as the middle frames of detected shots using TransNetv2 [22], [104], are utilized for further analysis. However, due to certain limitations, such as a lack of specialized features for AVS tasks and being operated by only one person, the AVS score achieved by diveXplore was relatively low. Another difficulty was the very coarse uniform frame sampling for the MVK dataset, which made many queries unsolvable.

Verge [84] incorporates a human and face detection module that accurately detects and counts human and human faces in the keyframes of each shot. The module uses the YOLOv4 deep neural network, which uses a DCNN architecture to extract human silhouettes and faces. The model is trained on the MS COCO dataset [69] and fine-tuned using the CrowdHuman dataset [102] to handle crowd-centered scenes with occlusions. During inference, the module calculates the total number of humans and human heads by considering only bounding boxes that surpass a predefined threshold. This enables users to distinguish activities involving single or multiple individuals effectively.

v-FIRST is among the first teams to use a prompt suggestor [59] powered by a Large Language Model (LLM) to suggest search terms based on the data to guide the retrieval process and enhance the clarity of the query. v-FIRST also suggests the adoption of external search engines to collectively form a system of specialized components.

G. BROWSING

Different VBS systems provide diverse browsing interfaces to facilitate the exploration of search results and enhance the user experience. Here, we summarize only their main features while directing interested readers to the official VBS webpage for short videos demonstrating each system's search and browsing capabilities (https://videobrowsershowdown.org/teams/vbs-2023-systems/).

vibro maintains its longstanding 2D map of visually arranged images, which has been improved in the 2023 version for faster and more accurate performance using Fast Linear Assignment Sorting (FLAS) [37] for grid layout arrangement and Dynamic Exploration Graph for internal graph representation [57]. These maps are mainly used to find similar images in the entire dataset, the current search result list, and the selected video. Another addition to vibro this year is automatic video playback in AVS mode when hovering over an image to quickly identify the motion of objects in the frame for more complex tasks.

In VISIONE's browsing interface, results are grouped by video, with each row representing one video and displaying up to 10 results sorted by the retrieval model's score. Each result has a menu that provides various options, such as performing similarity searches, viewing the temporal context, playing the entire video starting from the selected frame, or previewing the video in a neighborhood of the selected frame.

VIREO's browsing interface comprises three main components: a ranked list of video shots returned from the search engine, shots from the same video arranged chronologically when a shot is selected, and a pop-up window showing shots most similar to the selected one during similarity search.

vitrivr-VR is the only system with a virtual reality (VR) user interface, offering several result-browsing interfaces directly in VR. Query results are displayed cylindrically around the user, allowing intuitive browsing by turning the head. Additionally, the cylindrical grid can be rotated horizontally, hiding higher-ranked, already-viewed results and revealing unseen results in their place. For temporal queries, each space in the results grid shows a stack of previews, one frame for each segment in the matching result sequence. Results can also be grouped by video, such that each position in the grid shows the highest-ranked segments of a single video, ordered by the best-ranked segment. Intra-video browsing is facilitated by providing users with both a conventional video player with a timeline and a multimedia drawer showing keyframes of the video in a virtual box. By riffling through these keyframes, each can be intuitively inspected and selected to skip to the relevant part of the video.

The browsing interface of CVHunter consists of a scrolling grid of top-ranked video frames, with the option to quickly inspect video sequences (usually four fps) and apply presentation filters (i.e., select only k top-ranked frames from each video). From each frame, a video summary with representative frames is accessible.

vitrivr offers multiple result presentation options. In the context of VBS, the most relevant option groups retrieved segments by video and arranged them chronologically within each video. Each segment is shown using a static preview image. Clicking on any such preview opens a video player overlay that starts the playback of the video from the start of the selected segment. Additional controls, such as adjusting playback speed and navigating the timeline, facilitate browsing within the video.

Verge's user interface offers image size customization, an undo button for reverting to previous results, and a rerank button to rerank results based on another query. It offers several search modules, including free text search, concept and activity search, late fusion, temporal fusion, color-based image search, and search based on the number of people or faces. Recent enhancements in Verge include replacing the filmstrip of frames with a modal, a button for video playback directly on each shot, and a button that, when enabled, will return only the best shot from each video for AVS queries.

QIVISE's interface is organized into three distinct areas. After calculating the relevance of video shots, the system presents a sorted display in the main window. Users can click on any thumbnail to initiate one of two displays: the Video Shot Display or the Shot Segmentation Display. The Video Shot Display presents all shots from the same video as the selected shot. This is particularly useful for queries that span multiple shots, where a single shot may not provide sufficient information for accurate judgment. The Shot Segmentation Display, on the other hand, shows frames immediately before and after the chosen moment within the video, offering a more granular view of the surrounding frames to assist users in their evaluation.

VideoCLIP enables search using a variety of modalities, including rich text, dominant color, OCR, and query-by-image. The results are displayed in groups based on their video and video segments to reduce the effort for a user when locating potentially relevant targets.

v-FIRST supports browsing at different granularity levels to facilitate quick dataset browsing and discards similar images to enhance information density.

diveXplore [100] uses a simple 2D grid for browsing results. For each result, it is possible to inspect the context by opening the shot list of the corresponding video, which also provides a video player and displays available meta-data.

In 4MR's [34] browsing window, results are arranged in a grid, with the first 500 displayed. Each video segment is represented by its keyframe. Users can initiate a video preview by clicking on these keyframes, which opens a video player for further exploration.

Inspired by dating apps like *Tinder*,[5] Perfect Match's browsing interface presents a frame suggestion to the user based on the search input. The user can quickly decide if the suggested frame belongs to the desired shot. If it is correct, the frame can be submitted. Alternatively, the user can view the next frame suggestion or search for the same or another video.

IV. ANALYSIS OF VBS RESULTS

This section provides a comprehensive overview of the competition results and the performance of different systems.

[5]https://tinder.com

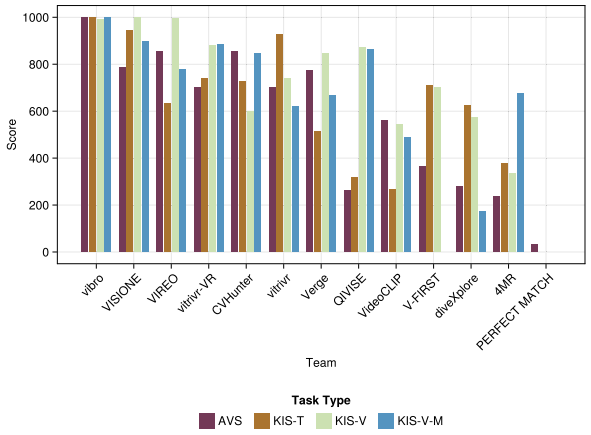


FIGURE 2. **Overall scores per team and task type.**

Subsection IV-A offers an overview of the overall results, while Subsection IV-B delves into the analysis of the KIS tasks, examining the available result logs for a select number of systems. Subsection IV-C provides information on performance in AVS tasks. The code and data required to replicate all the analyses presented in this section are publicly available via: https://github.com/sauterl/VBS23-Post-Hoc-Analysis.

A. OVERALL RESULTS

During the competition, a total of seven AVS, seven KIS-T, six KIS-V, and six KIS-V-M tasks were performed. The total normalized scores per task type and team are shown in Fig. 2. Among the four types of tasks, `vibro` achieved the best performance in three of them (AVS, KIS-T, and KIS-V-M), while `VISIONE` exhibited the best performance in the KIS-V task category. Therefore, those performances were awarded 1,000 points in their respective categories, and all other teams were scored proportionally. The overall score, as shown in Table 1, is the sum of these four scores per team.

Fig. 3a shows the number of correct and incorrect submissions per team for the three known-item search task types. `vibro` was the only team with seven correct submissions in the KIS-T category, making only two incorrect submissions along the way. For visual KIS tasks based on the V3C dataset, `VIREO` made the most correct submissions with six, followed by several other teams with five correct submissions each. However, `VISIONE` was the only team that managed to have five correct submissions without making an incorrect one. For the visual KIS tasks using the MVK dataset, `vibro`, `VISIONE`, and `vitrivr-VR` managed to make six correct submissions without any incorrect ones each.

The submissions manually judged for the AVS tasks are shown in Fig. 3b per team and status. `VIREO` made the most correct submissions, as well as most of the submissions in total. `vibro` made the second highest number of correct submissions across all tasks in this category but far fewer incorrect ones than `VIREO`, resulting in a higher total score.

The scoring function for the three types of KIS tasks considers not only the number of correct and incorrect

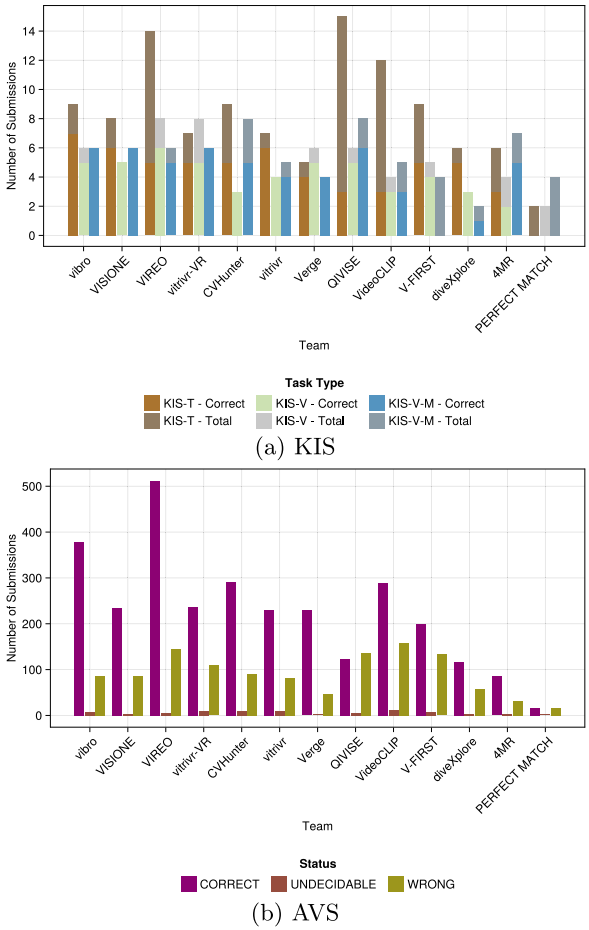
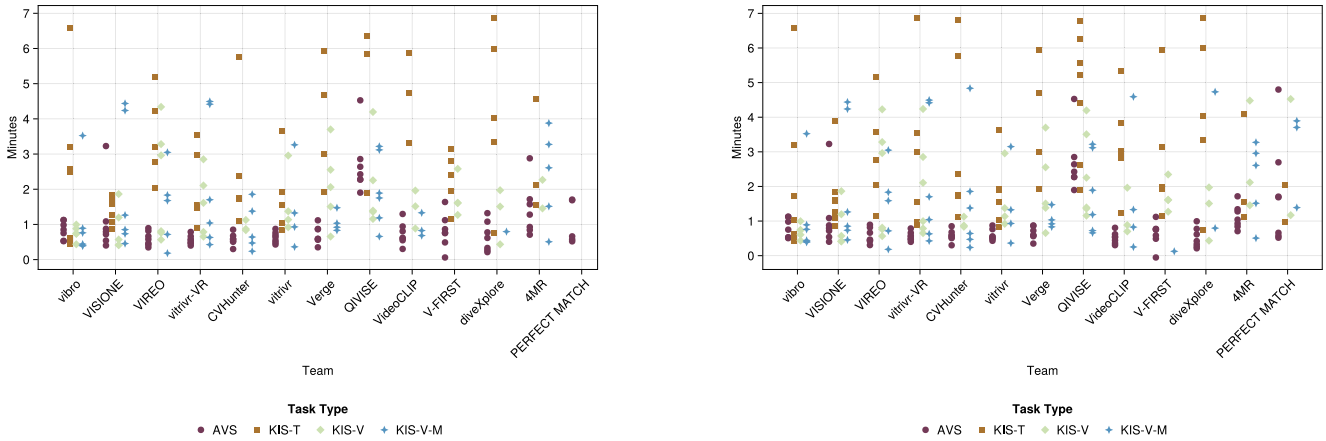


FIGURE 3. **Distribution of correct and incorrect submissions for KIS tasks (a), and correct, incorrect, and undecidable submissions for AVS tasks per team (b).**

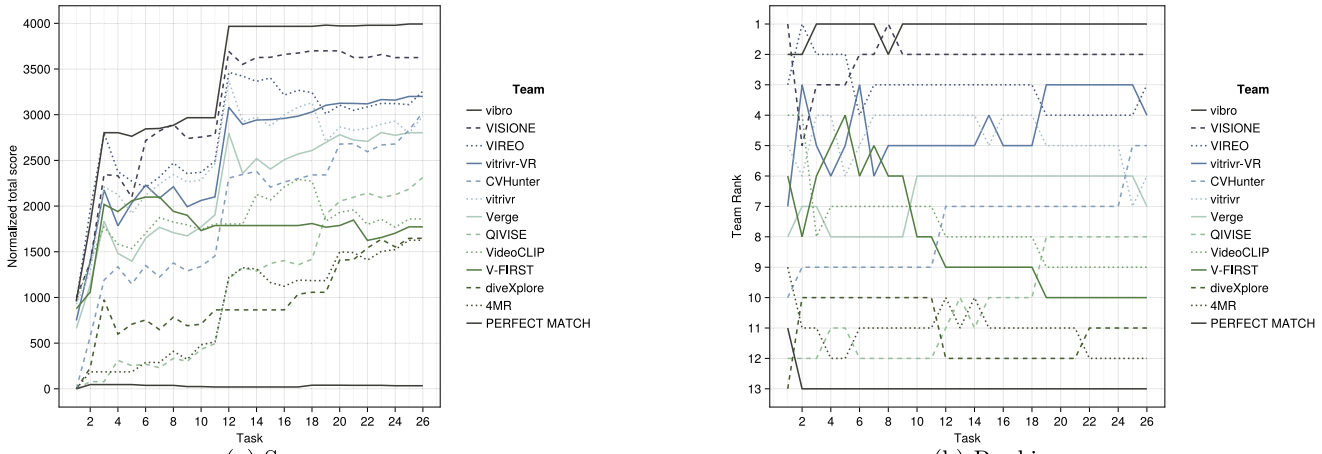
submissions per task but also the time in which the correct submission is made. Fig. 4a shows the time in minutes until the first correct submission was made, grouped by team and task type. In comparison, Fig. 4b shows the time until the first submission, regardless of its correctness. It can be observed that the teams with a higher total score not only managed to find the correct result for more tasks but also did so more quickly than others. This difference in time, for example, explains the difference in scores between `vibro` and `VISIONE` in the KIS-V-M tasks, despite both teams having the same number of correct submissions across all tasks, as can be seen in Fig. 3a.

The progression of the total normalized scores per team can be seen in Fig. 5a. Significant jumps occur every time a task from a new category (KIS-T, KIS-V, KIS-V-M, or AVS) is solved, as the scores are normalized to 1000 points within any category. The figure illustrates that while there are apparent differences in performance between certain teams, others performed very similarly and remained close in terms of total score throughout the entire evaluation. The evolution of the system ranks in the competition leaderboard, derived from these score developments, are shown in Fig. 5b. Interestingly, while the two highest-scoring teams changed places after the



(a) first correct submission time (b) first submission time

FIGURE 4. Distribution of time until the (first) correct submission per team and task type (a), and distribution of time until the first submission per team and task type (b).



(a) Scores (b) Rankings

FIGURE 5. Development of total normalized scores per team over time (a), and team ranking over time (b).

9th task and remained stable with respect to their ranking until the end of the evaluation, this is not the case for the teams with ranks 3 to 7.

B. ANALYSIS OF PERFORMACE IN KIS TASKS

In this section, we aim to gain deeper insights into how each system performed in KIS tasks by conducting an in-depth analysis of the logs collected by each team during the competition. These logs, structured in JSON format, contain essential information such as the team identifier, user identifier (when available), timestamp, query description, and a list of retrieved items for that particular query ordered by rank.

After a preliminary analysis of the available logs, it turned out that not all the teams had usable logs due to a heterogeneous set of problems (e.g., unrecoverable timestamps, incomplete records due to logging system failures, etc.). Teams with unrecoverable logs are excluded from the following analysis. Although many of the top systems correctly logged results for the tasks related to both the V3C and MVK datasets, the only one that presented dataset-specific logging problems was `Verge`, which had unrecoverable MVK logs. Despite this issue, we opted to include `Verge` in the analysis, resulting in a final pool of seven teams: `vibro`, `VISIONE`, `VIREO`, `vitrivr-VR`, `CVHunter`, `vitrivr`, and `Verge`. Notably, these seven systems also correspond to the top seven best-performing teams, according to the global competition leaderboard.

1) LOG PRE-PROCESSING

A thoughtful collection and analysis of the logs was performed to ensure that all the results of the different teams were comparable despite the strong heterogeneity of the logs.

Logs were retrieved directly from the DRES server for some systems like `Verge`, `VISIONE`, and `vitrivr`, while logs for other systems were obtained directly from the authors who saved them locally, including `CVHunter`, `vibro`, `VIREO`, and `vitrivr-VR`. The first step was to ensure

			V3C													MVK					
			T1	T2	T3	T4	T5	T6	T7	V1	V2	V3	V4	V5	V6	M1	M2	M3	M4	M5	M6
vibro	correct frame/shot	rank	1	1	12	13	4	16	-	1	660	54	10	22	12	8	4	6	5	41	3
		t_f	28s	6s	277s	140s	104s	21s	-	37s	17s	44s	21s	30s	280s	16s	30s	15s	94s	23s	11s
	correct video	rank	1	1	12	13	1	16	1	1	19	2	10	1	2	8	1	3	5	41	1
		t_v	28s	6s	277s	140s	79s	21s	23s	37s	29s	29s	21s	39s	297s	16s	30s	15s	94s	23s	11s
	correct submission	t_{cs}	37s	26s	395s	155s	150s	29s	193s	44s	53s	60s	26s	45s	-	27s	46s	24s	211s	53s	23s
VISIONE	correct frame/shot	rank	1	1	1	16	1	1	4	4	-	2	24	19	224	6	80	17	33	27	2
		t_f	26s	65s	384s	76s	88s	23s	64s	16s	-	26s	18s	62s	298s	23s	205s	12s	245s	19s	37s
	correct video	rank	1	1	1	16	1	1	4	1	106	1	24	1	2	4	80	17	33	27	1
		t_v	26s	59s	384s	76s	88s	23s	64s	16s	87s	26s	18s	65s	163s	23s	205s	12s	245s	19s	12s
	correct submission	t_{cs}	65s	76s	-	111s	100s	52s	95s	25s	112s	34s	25s	72s	-	76s	254s	27s	266s	51s	44s
VIREO	correct frame/shot	rank	6	141	100	20	201	5	-	185	-	40	141	-	-	132	78	216	188	198	5
		t_f	137s	409s	53s	187s	50s	24s	-	17s	-	13s	32s	-	-	16s	20s	16s	101s	219s	7s
	correct video	rank	1	141	64	20	6	5	1	50	77	1	16	6	20	26	78	40	188	198	4
		t_v	20s	409s	280s	187s	91s	24s	176s	22s	223s	13s	15s	36s	146s	16s	20s	72s	101s	219s	7s
	correct submission	t_{cs}	166s	-	-	253s	122s	311s	191s	49s	260s	34s	46s	178s	197s	43s	110s	101s	183s	-	11s
vitrivr-VR	correct frame/shot	rank	7	14	-	-	-	4	-	189	-	-	-	-	-	105	12	135	-	546	1
		t_f	27s	81s	-	-	-	34s	-	132s	-	-	-	-	-	13s	161s	40s	-	66s	16s
	correct video	rank	2	14	4	64	3	3	2	62	143	32	3	37	24	99	12	49	171	546	1
		t_v	85s	81s	40s	395s	130s	34s	39s	132s	39s	33s	119s	26s	83s	15s	161s	40s	215s	66s	16s
	correct submission	t_{cs}	213s	-	93s	-	179s	53s	88s	171s	97s	39s	126s	47s	-	38s	270s	62s	265s	102s	26s
CVHunter	correct frame/shot	rank	38	4	6	63	4	1	74	339	152	6	299	12	151	10	13	1	359	1	9
		t_f	268s	22s	317s	350s	89s	102s	38s	275s	96s	40s	57s	15s	34s	98s	67s	20s	241s	31s	10s
	correct video	rank	1	4	6	40	3	1	2	21	152	6	99	12	10	1	1	1	148	1	9
		t_v	268s	22s	317s	350s	95s	102s	51s	275s	96s	40s	57s	15s	34s	98s	67s	20s	255s	31s	10s
	correct submission	t_{cs}	345s	103s	-	-	106s	142s	66s	-	-	50s	-	52s	68s	111s	83s	28s	-	39s	14s
vitrivr	correct frame/shot	rank	4	4	2	289	1	81	22	65	-	40	43	165	31	200	24	6	-	16	1
		t_f	23s	78s	32s	88s	110s	26s	39s	20s	-	37s	135s	140s	43s	15s	201s	179s	-	38s	16s
	correct video	rank	2	4	2	7	1	81	1	30	202	1	43	12	31	46	24	6	641	16	1
		t_v	47s	78s	32s	88s	110s	26s	39s	20s	133s	31s	135s	140s	43s	21s	201s	179s	219s	38s	16s
	correct submission	t_{cs}	94s	219s	-	115s	116s	63s	50s	68s	-	55s	177s	-	82s	80s	-	196s	-	56s	22s
Verge	correct frame/shot	rank	3	79	638	132	39	3	-	518	-	393	82	122	9						
		t_f	40s	160s	289s	345s	271s	131s	-	201s	-	18s	55s	65s	204s						
	correct video	rank	3	79	124	132	14	3	25	17	86	2	82	122	2						
		t_v	40s	160s	210s	345s	369s	131s	27s	201s	33s	25s	55s	65s	120s						
	correct submission	t_{cs}	281s	-	-	356s	-	180s	115s	222s	124s	39s	153s	90s	-						

FIGURE 6. The table reports for each system with logs (i) the best-achieved rank of a *correct item* (frame or video shot); (ii) the time t_f in seconds when the best ranked correct item was retrieved; (iii) the best ranking of any frame/shot of the *correct video* (but not necessarily the correct video segment); (iv) the time t_v in seconds when the best-ranked video frame/shot was retrieved; (v) the time t_{cs} of the system's correct submission. Red values are for the best-detected ranks of the target video if the correct segment was not present in the logged result for a task. Green cells and Yellow cells show the best achieved correct item and video with a rank less than or equal to 10, respectively. Red cells indicates browsing failures when a correct item was in the first 1,000 results but was not submitted. Orange cells are other browsing failures when the correct video was present – but no correct frame or shot was present – and no correct submission was made.

that locally collected logs complied with the DRES format and that timestamps were consistent and synchronized with DRES local time. Records that did not fall into an active task were filtered out to ensure that only relevant actions were considered in the analysis. Due to factors beyond direct control, such as network problems or logging subsystem failures, a limited number of logs may be incomplete or not directly comparable.

It is important to note that different teams logged the retrieved results up to different maximum ranks. For example,

VISIONE, vibro, Verge, and CVHunter ranked the first 10K results, vitrivr and vitrivr-VR ranked the first 5K, and VIREO ranked the first 1K only. Moreover, for specific queries, the maximum rank may be even lower (for example, when computing the intersection between two result sets to work out temporal queries). Additionally, the time units logged vary between teams. In particular, for the V3C dataset, vibro, CVHunter, and VISIONE logged frames, while Verge, vitrivr, and VIREO logged segments (pre-defined shot IDs). For the MVK dataset, VIREO logged frames instead. To overcome this heterogeneity in logging units, we convert all temporal information into a physical-time format (seconds from the start of the video).

Despite these potential sources of errors, we consider that this level of uncertainty is sufficient to evaluate the team's browsing and retrieval capabilities. It is worth noting that during the competition, a live judge had the discretion to manually accept submissions from the same shot that fell just outside (less than 3 seconds) the KIS ground truth segment boundary. However, since such occurrences were rare, the original official ground truth was used for the subsequent analysis.

2) COMPARISON OF SYSTEM'S RETRIEVAL EFFECTIVENESS

The table in Fig. 6 presents the retrieval effectiveness of the various teams for both the V3C and MVK datasets, focusing on the best rank and best time at which the correct shot was found during the search. In particular, it includes the best-achieved rank of a correct item (either frame or shot) with the corresponding time indication (in seconds) from the start of the task, as well as the time of the correct submission.

This analysis was calculated not only at the level of the whole *team* but at the level of the specific *user* who used the tool since mixing the different users who used the tool may cause some unfair comparisons. Therefore, we report the results from the *best user* only, where the *best user* was identified as the one among the two that, for that particular task, obtained – ordered by decreasing importance – (i) the best shot rank, (ii) the best video rank, (iii) the shortest time when the best shot was retrieved, (iv) the shortest time when the best video was retrieved. Notice that we can distinguish between the two users from the team logs, but we miss the information about which user submitted the correct item. For this reason, we cannot define the best user as the one who submits the correct result, although it was a reasonable choice.

Examining the overall table, we observe that the first two teams, vibro and VISIONE, consistently achieved the best correct frame/shot within the first ten results in the majority of tasks (10 out of 19), and vibro achieved the best result on the challenging MVK dataset (in 5 out of 6 tasks). Furthermore, we notice a considerable variation among teams regarding the percentage of tasks for which the correct shot rank (or even video rank) is less than ten. Interestingly, this percentage does not always align with the ranking obtained in the final leaderboard. For example, although VIREO and vitrivr-VR (ranked 3rd and 4th, respectively) successfully retrieved the *correct video* within the first ten results in almost 40% of tasks, they achieved the *correct shot* within the first ten results less frequently than CVHunter or vitrivr (ranked 5th and 6th). This discrepancy might be attributed to two factors: (i) users manually searching for the correct shot within the correct video, allowing these systems to compensate for possible retrieval failures with effective browsing abilities, or (ii) interfaces that group results by video, enabling users to quickly locate the correct item within the first few videos with minimal scrolling.

The teams that experienced significant browsing failures, wherein the correct shot was present in the result set, but users were unable to locate and submit it within the allotted time, were CVHunter and Verge.

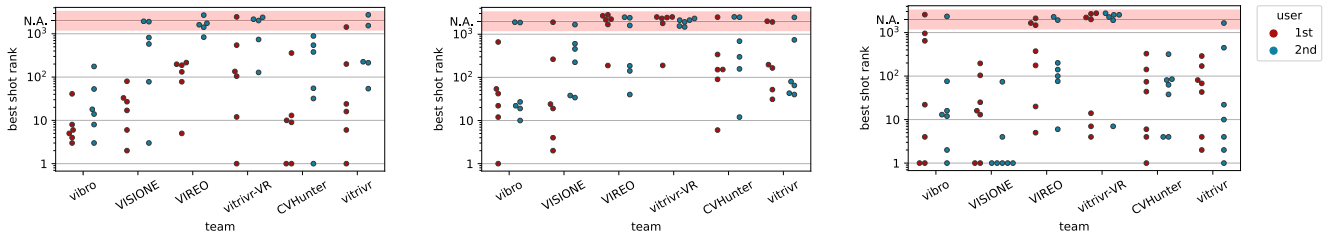
Despite the inherent challenges posed by the novel MVK dataset, characterized by highly redundant and noisy video content (involving moving cameras in underwater environments), all the teams demonstrated good performance, with a team-wise average percentage of incorrect submissions of 13% (only four incorrect submissions out of a total of 30).

In Fig. 7, we also report the best shot rank in the form of a scatterplot. Unlike the results reported in Fig. 6, we separated the two users and used the real user IDs instead of the calculated best and second best. This plot helps to understand if some users are noticeably better at querying their system. Note that Verge is not included in this figure due to limitations in its logs, which do not allow for an exact distinction between the two system users. In the V3C data set, the distributions of the best shot rank among the two users seem to intersect slightly for all the teams in the Textual KIS tasks, while more noticeable differences are visible in the Visual KIS tasks. We emphasize that this is just a hypothesis provided that the available data are limited by the VBS evaluation style (see Section V). On the MVK dataset, it appears that most teams have one user outperforming the other. This can be a direct consequence of the challenges introduced with the novel MVK dataset, which probably requires different searching and browsing strategies that are still not well established, therefore producing high variance inside the teams.

3) BROWSING EFFICIENCY

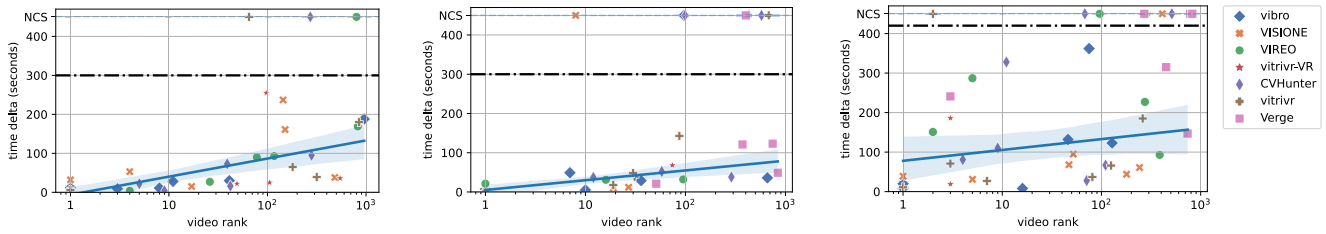
The time elapsed between the correct submission and the first appearance of a correct video in the logged result set is depicted in Fig. 8. We report the results for both V3C and MVK datasets, including the results from both users. It is important to note that these graphs provide an estimation of the actual browsing time, considering that a correct submission may have been made by inspecting the video rather than the top-ranked frames/shots. Additionally, the user who first retrieved a correct item may not be the same person who submitted the final correct answer, as this information is not always available.

Visual KIS tasks for both V3C and MVK datasets(in Fig. 8a and Fig. 8b) generally have low variance and a



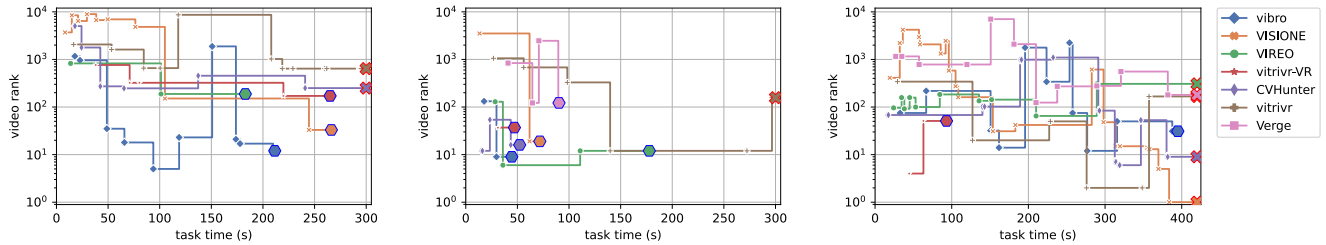
(a) Visual KIS (MVK) (b) Visual KIS (V3C) (c) Textual KIS (V3C)

FIGURE 7. Best rank of correct items appearing in result logs, for both MVK visual KIS tasks (a), V3C visual KIS tasks (b) and V3C textual KIS tasks (c).



(a) Visual KIS (MVK) (b) Visual KIS (V3C) (c) Textual KIS (V3C)

FIGURE 8. Relation between the rank of the first occurrence of a video in the result logs and time delta to correct submission, for visual KIS on MVK (c), and visual (a) and textual (b) KIS on the V3C dataset. The black dash-dotted line represents the duration of the task. *NCS* stands for *Non-Correct Submissions* and corresponds to all the correct frames found in the result logs that were not submitted correctly (either due to running out of time or incorrect submissions). The blue line is found by linear regression and is accompanied by the 95% confidence intervals.



(a) Visual KIS (MVK) - M4 (b) Visual KIS (V3C) - V5 (c) Textual KIS (V3C) - T3

FIGURE 9. Browsing storyline for some of the KIS tasks. We report the browsing storyline only for the best user, indicating the correct submission with big hexagons and the wrong submissions with crosses.

slight slope. The low variance of the fit indicates an effective positive correlation between the rank of the initial result and the time of correct submission. Additionally, the slight slope suggests that for Visual KIS tasks, most of the users were able to submit the correct result even if the first occurrence of the shot/frame in the logs had a high rank.

A different scenario occurs with the Textual KIS tasks in Fig. 8c, where there are cases in which the correct item is found within the first ten results, but the users cannot find it within the first 100 seconds, and in a particular case, `vitrivr` cannot correctly submit it. This evidences the intrinsic difficulty in Textual KIS tasks, where the semantic gap between texts and images makes searching and browsing more challenging.

In Fig. 9, we also report the temporal evolution of the video rank for all the teams throughout particular tasks. Note that these plots show changes in rank only when information appears in the logs. Therefore, the flat line between two consecutive changes in rank is a loose representation of reality. The rank is also changing due to some unreported browsing actions performed by the team during the "flat line" zone. Despite these shortages, these graphs provide valuable insight into the results already reported in Fig. 6. For example, while there are tasks in which most of the teams retrieve results after the first query formulation (Fig. 9b), there exist other more complex tasks (like the KIS-T task in Fig. 9c) in which the video rank oscillates broadly until some teams can retrieve the correct video within the first ten results after more than 250 seconds. At this point, due to lack of time, most teams fail to submit correctly, with `vitrivr` and `VISIONE` failing after finding the correct video in the 2nd and 1st positions, respectively. These plots strongly underline the importance of interactive search for solving the proposed tasks.

4) ANALYSIS OF USER QUERIES

In this section, we conduct a detailed examination of the types of search queries formulated during the competition and

TABLE 3. Percentage of how many queries were text and how many were other query types. The category "Text+Temporal" indicates when two or more text queries are used together to perform a temporal search.

	query type usage		
Team	Text	Text+Temporal	Other queries
vibro	46.6%	6.8%	46.6%
VISIONE	58.3%	38.0%	3.7%
vitrivr-VR	94.5%	5.5%	0.0%
CVHunter	54.4%	10.1%	35.5%
Verge	88.3%	0.0%	11.7%

assess their level of success. Our specific focus is directed toward the integration of joint text-image embeddings and how these embeddings evolve with each reformulation of the text queries.

For these investigations, it is imperative to have access to log data that includes both the query type and the actual formulated query. Among the teams examined, only the log data from vibro, VISIONE, vitrivr-VR, CVHunter, and Verge are considered suitable for our research. Consequently, this section exclusively emphasizes these five teams. It should be noted that users associated with the team Verge submitted their data using identical user IDs, making it unfeasible to distinguish between user 1 and user 2 in subsequent analyses. Additionally, in our analysis, we removed duplicate queries from the same team and user when these duplicates were caused by logging problems or when a user had submitted the same search query multiple times.

Table 3 compares the most frequently used query types. It becomes evident that most teams have primarily employed text-based searches using joint embeddings. Other queries that are not text-based are often solely image-to-image searches, except for a filtering query by Verge. CVHunter and vibro are the only two teams that utilize content-based image retrieval more frequently, with vibro achieving an almost 50:50 ratio. Regarding text queries, VISIONE was the team that used them most often in combination with temporal searches.

In Table 4, we compare the average performance of text queries, text queries used in combination with temporal search, and all other query types per user. The user column represents the first and second users in each team's log files. The Top-k columns indicate the percentage of queries for which the target shot appeared in the Top-k results. A dash ("-") denotes that no searches of the respective query type were conducted. The "query / min" column represents the mean number of queries conducted per minute for a particular query type and user. Since a task is considered completed for a team as soon as one of the two team users submits a correct result, we computed the individual user's queries per minute for a given task by dividing the time it took the team to submit the correct result (or the total duration of the task if unsolved) by the number of queries made by that individual user within that time frame. Additionally, the average word count and character count (query length) are provided for text queries.

When analyzing text-based queries only, there are no significant differences between team users in terms of queries per minute. Deviations are mainly due to longer text queries, more frequent use of temporal queries (2nd user of VISIONE), or, in the case of CVHunter, the utilization of other query types. The two users within a system seem to respond equally quickly, except the second user from VISIONE, triggering twice as many queries as their first user. This discrepancy might arise from the possibility of sending multiple similar queries (such as correcting spelling or punctuation errors, to which text-to-image-based embeddings are particularly sensitive). The relatively low number of queries in vitrivr-VR can be attributed to the slower process of inputting queries in the virtual reality user interface, as the system primarily emphasizes browsing functionality.

Furthermore, there appears to be a slight correlation between the number of words per query or the query length and the ranking of the target shot. In each team, the user with longer search queries achieved better Top-k rankings. This can be attributed to CLIP's capability to process detailed written information, thus enhancing the search results. The results are not comparable between teams due to the use of different CLIP models. Nevertheless, a more in-depth analysis of the relationship between query length and the ranking of correct results for both users within a team indicates that there isn't a clear-cut correlation. In general, for VISIONE and CVHunter, longer queries tend to achieve better rankings for correct shots than shorter queries. However, systems such as vibro, vitrivr-VR, and Verge exhibit a more balanced distribution in this regard.

The Top-100 values for the type of text query among different users in teams vibro and VISIONE are quite similar, suggesting that their text search abilities are also fairly comparable. vibro's slightly superior ranking in the competition could potentially be attributed to the application of other query types and browsing functionalities.

For the "Other" query type rows, only vibro and CVHunter can be considered, as other teams rarely employed anything other than text queries, and their Top-k values for other query types might contain outliers. Both vibro and CVHunter use search-by-example queries. Although their average results for Top-10, Top-20, and Top-50 are inferior to their text results, the figures for Top-100 and beyond are relatively similar. Therefore, both systems demonstrate the ability to perform context-based image searches that effectively complement text-based queries. Please note that there is a dependency between the rows corresponding to different query types, as simple tasks are usually solved by an initial text query, whereas other query types are commonly employed in subsequent steps to address more complex tasks.

TABLE 4. Query statistics per team member and query type averaged over all KIS tasks. The queries per minute, average number of words, and string length of textual queries are depicted for each user. Additionally, top-*k* denotes the percentage of queries for which the target shot was within the first *k* results.

team	user	query type	# queries (usage)	query per min	words per query	query length	top10	top20	top50	top100	top200
vibro	1st	Text+Temporal	3 (4.1%)	0.1	13.0	61.0	0%	66.7%	100.0%	100.0%	100.0%
		Text	34 (45.9%)	1.3	9.7	51.0	14.7%	17.6%	29.4%	35.3%	44.1%
		Other	37 (50.0%)	1.4	-	-	5.4%	13.5%	27.0%	35.1%	37.8%
	2nd	Text+Temporal	8 (9.2%)	0.3	8.0	43.0	0%	0%	0%	0%	0%
		Text	41 (47.1%)	1.5	7.0	37.3	12.2%	22.0%	29.3%	34.1%	39.0%
		Other	38 (43.7%)	1.4	-	-	2.6%	10.5%	21.1%	28.9%	36.8%
VISIONE	1st	Text+Temporal	10 (21.3%)	0.4	32.8	173.9	10.0%	30.0%	60.0%	80.0%	90.0%
		Text	37 (78.7%)	1.5	17.1	87.9	10.8%	13.5%	24.3%	29.7%	37.8%
		Other	-	-	-	-	-	-	-	-	-
	2nd	Text+Temporal	52 (45.2%)	2.1	19.6	98.6	13.5%	17.3%	25.0%	30.8%	46.2%
		Text	58 (50.4%)	2.3	10.3	52.9	10.3%	17.2%	22.4%	25.9%	32.8%
		Other	5 (4.3%)	0.2	-	-	20.0%	20.0%	40.0%	40.0%	60.0%
vitrivr-VR	1st	Text+Temporal	2 (3.6%)	0.1	14.5	78.5	0%	0%	0%	0%	0%
		Text	54 (96.4%)	1.7	7.1	34.9	7.4%	13.0%	14.8%	18.5%	22.2%
		Other	-	-	-	-	-	-	-	-	-
	2nd	Text+Temporal	3 (8.6%)	0.1	8.0	40.0	0%	0%	0%	0%	0%
		Text	32 (91.4%)	1.0	5.2	27.9	3.1%	3.1%	3.1%	3.1%	12.5%
		Other	-	-	-	-	-	-	-	-	-
CVHunter	1st	Text+Temporal	22 (21.2%)	1.1	15.2	79.7	31.8%	36.4%	40.9%	45.5%	50.0%
		Text	54 (51.9%)	2.7	10.4	57.4	11.1%	11.1%	13.0%	20.4%	29.6%
		Other	28 (26.9%)	1.4	-	-	10.7%	10.7%	14.3%	25.0%	35.7%
	2nd	Text+Temporal	1 (0.8%)	0.0	9.0	46.0	0%	0%	0%	0%	0%
		Text	70 (56.5%)	3.5	7.0	40.6	4.3%	5.7%	8.6%	12.9%	24.3%
		Other	53 (42.7%)	2.6	-	-	1.9%	7.5%	13.2%	18.9%	20.8%
Verge	1st & 2nd	Text+Temporal	-	-	-	-	-	-	-	-	-
		Text	109 (85.8%)	3.9	5.0	28.3	0.9%	1.8%	3.7%	5.5%	30.3%
		Other	18 (14.2%)	0.6	-	-	11.1%	11.1%	11.1%	16.7%	16.7%

In the following, we take a closer look at the joint text-image embeddings of the text queries used by all five teams, each employing some form of CLIP embedding. While the specific type of embedding may vary among the teams (refer to Section III-A), these embeddings were not stored in the logs. To facilitate comparison, we uniformly generated new embeddings using vibro's feature vector pipeline, as detailed in Section III-A. Then, we focused on the cosine distances between embeddings of individual queries provided for the same task.

Fig. 10 shows the mean pairwise cosine distances separately for each task and each team, as well as aggregated for task types and teams.[6] Several interesting observations can be made from Fig. 10. Firstly, the mean embedding distances are smaller for textual KIS tasks than for visual ones, especially V3C datasets. We attribute this difference to the fact that in textual KIS tasks, the scene description serves as a central point from which all users start developing their queries. In contrast, in visual tasks, many possible visual cues are causing higher variance in their respective textual descriptions. Furthermore, the higher variance within the searched scenes in V3C compared to MVK could further explain the difference between the two datasets. Secondly, there is a difference in the mean query distances per team. On average, from the least to the most diverse queries, the ordering is VISIONE (0.47), vibro (0.50), CVHunter (0.56), Verge (0.64), and vitrivr-VR (0.68).[7] The same order of teams is also achieved if we consider the queries of individual users separately. Note the correspondence between decreasing query distances and increasing overall team scores. We can assume that the longer the task remains unsolved, the more distant queries are produced by the team members.

To corroborate this hypothesis, we focused on the sequences of text queries constructed by individual users

[6]We only included points where at least two queries per team were available, and each pair of queries was considered with equal weight when calculating averages.

[7]Note that (i) both vitrivr-VR and Verge had missing data for some tasks, (ii) both VISIONE and vibro had significantly more compact queries than CVHunter on average (t-test p-value < 2.4e-6).

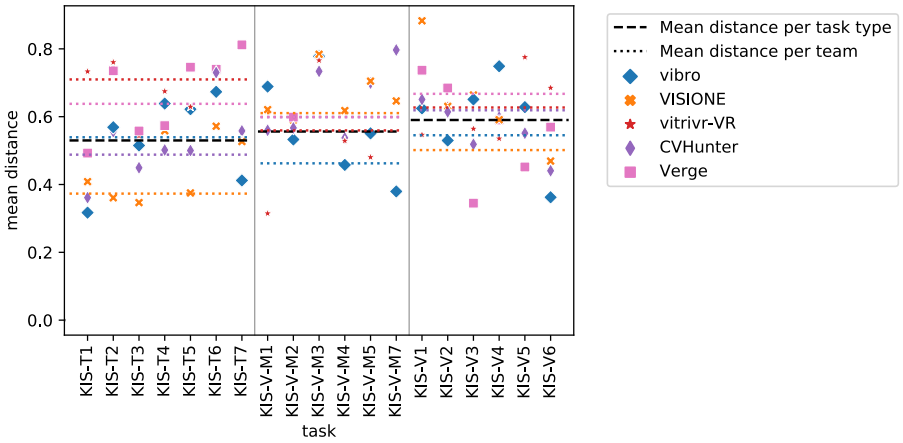


FIGURE 10. Mean pairwise query distances w.r.t. CLIP embedding features and cosine distance for individual teams and tasks. The average distances for each team and task type are depicted as dotted lines, while average distances w.r.t. all teams for specific task types are depicted as dashed lines.

TABLE 5. Comparison of mean distances in sequences of text queries. For both the distance w.r.t. CLIP embeddings (CLIP) and Levenshtein distance. The first line depicts distances from the first query, while the second line depicts distances to the previous query.

		Q2	Q3	Q4	Q5
CLIP	first query	0.41	0.53	0.57	0.58
	previous query	0.41	0.40	0.34	0.33
Levenshtein	first query	0.26	0.38	0.42	0.47
	previous query	0.26	0.24	0.18	0.19

for each task.[8] Table 5 presents the mean distances of the first text query provided by a user to a given task and the subsequent ones (i.e., Q1 vs. Q2, Q1 vs. Q3, etc.) as well as the distance between each query and the previous one (i.e., Q1 vs. Q2, Q2 vs. Q3, etc.). It is apparent that while the distances between subsequent queries remain roughly the same (or slightly decrease), the distance from the first query gradually increases. To verify these findings, we also conducted the analysis using the Levenshtein distance between query strings, yielding similar observations. The current data show that the distance to the initial query could converge[9] around the fifth query, but additional data with longer sequences would be necessary to verify this assumption.

Finally, we focus on what is the source of diversity in per-team text queries. For this, we compared the mean differences of queries for the same task within each user, between queries of both users from the same team, and between queries of users from different teams. The mean distance between queries of the same user was 0.46, the mean distance between queries of different users of the same team was 0.61, and the mean distance between queries of users of different teams was 0.64. We also checked for differences between individual teams, but no notable exceptions appeared. We interpret these results as follows. Individual users tend to be consistent in how they construct queries throughout the search task, with minimal variation. While the search tool itself also seems to play some role in the inter-query differences, users themselves (even those from the same team) are the primary source of diversity in query construction. This finding may support the argument for modifying the VBS competition to ensure more uniform user sampling.

[8] In the subsequent analysis, we only considered sequences of five or more queries. We removed the results of the `Verge` team as we could not reliably identify individual users in their logs.

[9] I.e., subsequent queries on average are not more distant from the first one than the previous ones.

TABLE 6. AVS tasks conducted during VBS 2023.

Task	Hint
a1	Find indoor shots of three or more people sitting around the same table with food on it.
a2	Find shots of only one person riding a horse or riding a horse-drawn cart, without other people visible.
a3	Find shots of an adult person running in an (sub-)urban street. Cars may be visible, but no other people are walking or running.
a4	Find shots of one or more adult persons holding, releasing, throwing, or playing with a balloon of any shape. Balls of any kind are not balloons; air balloons are not included.
a5	Find shots showing at least one person singing and at least one drummer (not necessarily playing the drums at that moment).
a6	Find shots taken by a paraglider or parachutist, where their shadow (the glider and/or the person) is visible on the ground or water.
a7	Find shots showing a bar chart (vertical or horizontal bars).

C. ANALYSIS OF PERFORMANCE IN AVS TASKS

In this section, we delve into the setup, evaluation, and analysis of the AVS tasks. Unlike the KIS task, the correct

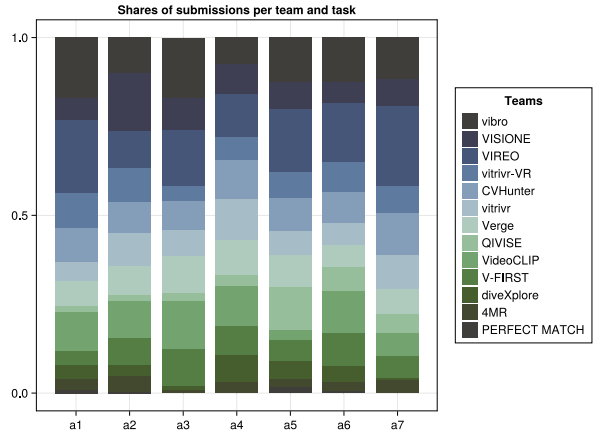


FIGURE 11. AVS: Ratios of the teams' first correct submission per task. When multiple teams submit the same segment, this contributes to each team's ratio individually.

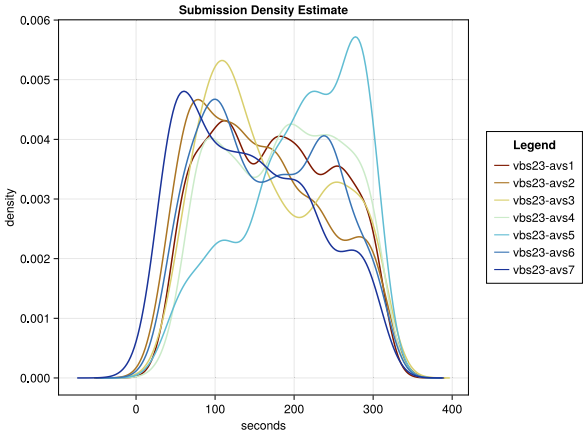


FIGURE 12. Kernel density estimate for AVS submissions. When multiple teams submit the same segment, this contributes individually.

answer to an AVS task is not unique. Instead, participants are tasked with finding as many relevant results as possible based on a brief textual description (see Table 6). Due to the massive dataset volume, it is unfeasible to label everything for ground truth. Therefore, real-time evaluation during the competition is conducted by experienced judges using the DRES evaluation server. Since there is heterogeneity in the submission units used by different teams (e.g., single frame number, specific time of the video, or a pre-defined video shot identifier), in DRES each submission is mapped onto pre-defined reference shots that will be presented to the judges for their evaluation. As detailed in Section II, penalties are applied for incorrect submissions to prevent excessive arbitrary submissions. Additionally, the evaluation metric accounts for diversity, as submitting different correct shots of the same video does not increase a team's score. Moreover, a video submission that is distinct from other teams' submissions can have a greater impact on the evaluation compared to a video that is commonly found by most teams. Specifically, as shown in Eq. (1), each team's score is divided by the total count of correct videos among all teams' submissions.

Fig. 11 compares the share of the first correct submissions from different teams in each task. We only count the ratio of the first correct submission, as it contributes the most to the AVS score (see Eq. (1)). Investigations have shown that even though the newly introduced scoring function results in diminishing returns, teams have submitted from the same video. However, since the scoring respects the diversity of videos found per team, the submission of the same video by multiple teams contributes independently to their shares. Each team did not consistently submit a similar share in different AVS tasks. For example, although VIREO tends to have the highest number of videos in most tasks, its number of submissions is significantly less in task a2. A similar situation can also be observed for vibro in task a2 and a4. Conversely, VISIONE dominates the share in task a2, despite having a relatively lower quantity of submissions in the other tasks. This fluctuation in a team's submission share can be attributed to various factors such as the features utilized, methodologies for composing queries, and search strategies employed. Furthermore, the teams with the highest number of video submissions in the AVS tasks, namely VIREO and vibro, achieved the top two highest AVS scores. This underscores the significance of the number of submissions in the evaluation process. Furthermore, although VIREO submits the most videos, its overall score is less than vibro as they submit more incorrect results, which can be observed in Fig. 3b.

In addition to the variation in submission density among different teams, submissions are not evenly distributed across the time dimension. Fig. 12 shows the estimated submission density during the competition time. In most tasks, the submissions reach their peak around 100 seconds, indicating that participants have typically formulated their queries and submitted the top-ranked results by this time. As the competition progresses, while participants continue to find new results, the density of submissions decreases. After the peak submission period, participants often need to revise their query, change their search strategy, and delve deeper into the ranked list to uncover additional results. In some cases, the submission peak occurs later in the competition. For example, in task a5, it takes nearly 200 seconds longer to achieve the submission peak, reflecting the greater challenge in formulating effective queries to optimize the ranking of matched videos.

The difficulty of an AVS task is reflected not only by the hardness of composing a suitable query but also by the difficulty of understanding the query and discerning the results. Fig. 13 presents two metrics to assess task difficulty. The bar chart shows the ratio of correct and incorrect submissions in each AVS task, with the x-axis indicating the tasks sorted in increasing correctness ratio. A lower correct ratio is mainly attributed to indistinguishable results. In addition, as is shown in Eq. (1), if all submissions

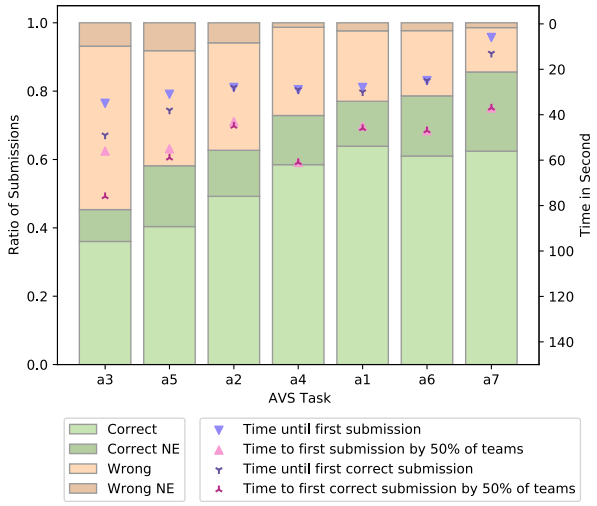


FIGURE 13. Ratios of the overall correct and wrong submissions per task (bar chart, left y-axis) and the time cost to submit the first (correct) result (scatter chart, right y-axis). Since only the first correct submission from a video is evaluated, both the correct and wrong submissions were divided into valid and not-evaluated (NE) submissions. The queries in the x-axis are in the increasing order of the correct submission ratio.

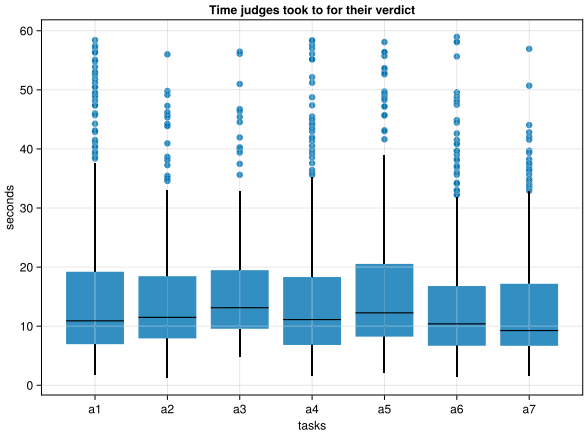


FIGURE 14. AVS: Time judges needed to render a verdict.

from a team referring to the same video are incorrect, they are all evaluated (with a penalty) in the final score, but if at least one correct submission exists, then only the submissions before the first correct submission from that video are evaluated. Therefore, we categorized both correct and wrong submissions into evaluated and not evaluated (NE) submissions to showcase the ratio of valid submissions. This categorization reveals that the majority of submissions come from distinct videos, while between 17% to 30% of correct submissions and 5% to 20% of incorrect submissions are not evaluated across queries. The difficulty in solving task a2 and task a3 arises from quantity and negative constraints, as shown in Table 6. The negative constraint "without other people visible" (a2) and "no other persons walking or running" (a3), and the quantity constraint, "only one person" (a2), require participants to check the entire video segment and perform the identification painstakingly. Participants are likely to submit a video segment solely according to its keyframe while overlooking the incorrect frames in the rest of the video segment. On the other hand, the scatter chart presents the time until the first (correct) submission by the first and 50% of the teams. This metric reflects the difficulty in understanding the query, composing a draft query, and submitting the first (correct) result. The y-axis on the right-hand side of Fig. 13 indicates the corresponding time in seconds. The lower the position of a scatter point, the more time is spent. The trends of difficulty level by the two different metrics agree with each other. An exception occurs in task a4, where the longest query length and the most complicated query structure (in terms of the depth of the dependency tree) lead to more time for 50% of teams to submit their first (correct) result, although the results are relatively easy to distinguish.

Fig. 14 presents the distribution of judgment time for each submission in different tasks. The median time for an experienced judge to give a verdict is around 10 seconds, which is apparently longer than the time for a participant to submit a video segment. Upon filtering outliers, there is no significant variation in judging time among different tasks. Nevertheless, we can still find that the trend aligns with the difficulty observed in Fig. 13 when sorting the time to judge in increasing order, indicating that the difficulty of submitting a video correlates with the difficulty of rendering a verdict at the task level.

Table 7 shows the agreement and disagreement between the judges and the teams. Each cell in the table represents the fraction $\frac{\#agreement}{\#disagreement}$, where $\#agreement$ and $\#disagreement$ represent the number of identical submitted shots judged as correct and wrong, respectively. For instance, in task a2, the red cell showing 0/1 indicates that while no shot was submitted by seven teams and judged as correct, there was one shot (item 02964, timed from 226 to 230 seconds) that was submitted by 7 teams and evaluated as incorrect. This discrepancy may arise because judges assess the entire shot, while teams may have viewed or submitted only a specific frame of the shot. For example, in the case of video 02964, while half of the shot is correct (showing only one person riding a horse), the other half contains several people riding horses, rendering it incorrect. Compared to last year [70], significant disagreement appears in more tasks, i.e., task a2 and a3. In task a2, seven teams disagreed with the judgment on one video, while four teams disagreed on one video in task a3. These tasks exhibit greater submission difficulty in Fig.13 and judgment difficulty in Fig.14, which could contribute to the higher disagreement. Beyond that, the disagreement is not significant in all other AVS tasks. On the other hand, we could see that if the description has a very low potential for misinterpretation (task a7), the majority of submissions are correct and in agreement with the judges' evaluation.

Regarding the newly introduced scoring function, its overall scores' rank correlation to the ranks obtained using the old formula is very high (0.929). Our interpretation of this

TABLE 7. Number of submissions in agreement/disagreement between judges and different number of teams.

Task	Number of Teams												
	1	2	3	4	5	6	7	8	9	10	11	12	13
a1	422/180	108/22	30/2	11/1	1/0	1/0	-	-	-	-	-	-	-
a2	112/117	28/19	21/7	12/0	3/0	3/1	0/1	1/0	-	-	-	-	-
a3	76/138	23/11	6/4	0/1	1/0	-	-	-	-	-	-	-	-
a4	200/139	54/10	24/2	8/0	5/0	1/0	-	-	-	-	-	-	-
a5	262/195	7/2	-	-	-	-	-	-	-	-	-	-	-
a6	231/120	58/5	17/0	10/0	4/0	-	2/0	-	-	-	-	-	-
a7	127/72	45/2	33/0	15/0	4/0	4/1	2/0	2/0	2/0	2/0	-	-	-

Red font highlights cases where the fraction is lower than or equal to one (i.e., $\frac{\#\text{agreement}}{\#\text{disagreement}} \leq 1$)

value is two-fold: a) there would not have been a significantly different ranking of the teams using the old formula, and b) the teams' search strategy appears unaffected by the scoring. However, since there is no survey on the strategy, we can only assume the latter point as the communication of the organizers during the competition has been for 2022 and 2023 to "find as many shots as possible".

V. CONSIDERATIONS FOR FUTURE EVALUATION SETUP

In this section, we suggest re-evaluating the VBS setting to address certain inherent limitations in its assessment methodology, which are detailed and discussed below. Specifically, we compare three alternative options:

- Collaborative Users in a Single Team (VBS 2023 Setting): In this setup, users of the same system are treated as a single team allowed to collaborate, and their cooperation contributes to the evaluation scores. A KIS task is considered completed for a team when the fastest user in the team submits a correct result, while incorrect submissions made by other team members incur penalties affecting the final score. In AVS, submissions from all team members are combined and scored collectively, with only one correct result evaluated per video if multiple users submit results from the same video. The winning system is determined by the team with the highest cooperative score.
- Independent Users Aggregated into a Single Team: In this configuration, independent users of the same system are considered a unified team, but direct collaboration among them is not allowed. Each user independently attempts to solve tasks, and the final score for a task is calculated as the average of individual user scores. The winning system is the team with the highest score.
- Independent Users as a Distinct Team: In this scenario, each user of a system is treated as a distinct team, with no collaboration or score aggregation among independent users. The VBS competition is won by the system with the highest-performing independent user.

The primary drawback of the VBS 2023 Setting is that once one user solves a task, the entire team stops searching, resulting in the loss of valuable information for analysis. Moreover, it is unclear if there is an outstanding user within a team, as individual performances are not discernible. Additionally, while this setting allows for the evaluation of cooperative systems, it puts systems that participate with a single user at a disadvantage.

The second and third options share two main advantages: Requiring all users to attempt to solve all tasks could enhance the competition's entertainment value and provide more comprehensive system evaluations. The second option may seem promising if all teams could send the same number of users to VBS. However, ensuring equal participation is challenging and cannot be guaranteed (e.g., financial constraints, staffing limitations, or other practical considerations). Moreover, some teams may choose to participate with only their best user (super-user), potentially limiting the volume of log data available for analysis. However, there are several compelling arguments for the third option:

- Any number of users (with a recommended minimum of 2) can participate without notable unfair effects on the overall system score, allowing for more extensive data collection for scientific analysis.
- Teams are motivated to have more users, as the performance of the best user is unaffected by additional team members. Moreover, it is always advantageous for a team to participate with as many users as possible as designating one system user as the "best" before the competition does not necessarily guarantee that they will ultimately deliver the best performance.
- Having independent users makes it feasible to examine whether there is a low or high variance in the performance of different users within the same system, which might not be possible in the other two scenarios analyzed, both of which involve aggregated scores.
- Identifying and analyzing super-users is more effective, as their impact on rankings can be assessed more clearly.
- There is no need to update the current infrastructure (DRES), though some aggregated visualization could be added for clarity.

Considering these factors, we recommended adopting the third option, which has been approved by the VBS organizing

committee for the 2024 edition of the competition. There are other aspects, however, that require attention. Improving the clarity of AVS queries and ensuring consistent logging practices are essential.

Our evaluation revealed that there is often a lack of agreement regarding the correctness of AVS submissions. Teams often seem to disagree with judges, and also judges sometimes disagree with each other. For example, we could see that the AVS task a3 was misunderstood by many teams (four teams submitted the same shot, which was rated as wrong) because it was probably not entirely clear what "*a sub-urban street*" is, and whether "*other visible people where really standing or running*" (or whether a marathon participant taking a short break is still "running"). This is very unfortunate for teams because incorrect submissions have a severe impact on the scoring. Clear task descriptions with minimal chance of misinterpretation, accompanied by visual examples, can reduce disagreements between teams and judges.

Another consideration for future VBS editions is to ensure that all participants adhere to comprehensive and consistent logging practices. Participants should be explicitly instructed to meticulously save logs of their systems, encompassing snapshots of results of all performed queries (at least top-k items), including query specifications, a practice already adopted by some teams. Including browsing actions in logs could further enhance understanding of users' interactions. However, non-trivial open challenges need to be resolved first [89].

Additionally, future evaluations should consider collecting specifications of participating systems to highlight the diverse hardware and computational resources utilized. This data would offer valuable insights into the performance and capabilities of these systems, enabling an analysis of the trade-off between efficiency and effectiveness.

Finally, it is worth noting that before the COVID-19 pandemic, the competition also included evaluation sessions with novice users. Unfortunately, these sessions have been omitted in recent years due to logistical constraints from remote participation. It is of great importance to make sure that the next editions of VBS include evaluation sessions with novices to provide a more complete evaluation of the usability and performance across diverse users.

VI. DISCUSSION AND FUTURE CHALLENGES

Our analysis indicates that teams predominantly relied on free-text search with joint embeddings, such as those derived from models like CLIP and OpenCLIP, complemented by result browsing. This approach proved effective for most KIS tasks. However, while visual KIS tasks were relatively easy with this type of search, textual KIS tasks posed greater challenges. Teams often struggled to formulate text queries that returned relevant results, with some tasks requiring more than 250 seconds to find a correct answer – this is clear evidence that content-based search still suffers from semantic gap when no visual example is available.

The observed variability in search performance among teams underscores the need for continued exploration of diverse search strategies and methodologies. For instance, the best two teams, `vibro` and `VISIONE` were able to find the search item within the first ten results for 10 out of 19 tasks, while other systems often ranked the correct item much higher, even though almost all systems shared the use of the latest CLIP models. This underscored the importance of complementing CLIP-based cross-modal search features with other effective search and browsing functionalities. For example, `VISIONE` stood out for its frequent use of temporal queries to complement textual queries, while `vibro` often complemented its textual queries with query-by-visual-examples. A promising direction is to support bi-modal queries, where visual and textual queries are combined to give the user fine-grained control over the properties of the target item. For example, the emerging field of Composed Image Retrieval [29] addresses the problem of retrieving target images that are visually similar to a query image but with modifications indicated by a textual query.

Another finding is that longer text queries often return more accurate results and that teams try to adapt their queries until some relevant content is presented in the top-k results. Interestingly, the diversity of free-text queries is relatively low within the same user but higher among different teams or even different users within the same team. While different understandings and formulations of the query are primary reasons for this, other factors, such as cultural differences, may also influence query diversity. Understanding the dynamics of user interaction with search systems, particularly the iterative adaptation of queries, can provide important insights into developing more intuitive and user-friendly interfaces (e.g., integrating automatic query suggestions). `v-FIRST` has been a pioneer in this regard, being the only VBS system that integrates an LMM-based suggestion tool to enhance query clarity. Although its performance did not excel in terms of competition ranking, the direction taken is promising. Recent advances in the field of LMMs, with continually improving performance, suggest that such models have significant potential in enhancing interactive video search.

There were also significant differences in the number of issued queries by the team and team members. VR systems seem to have some drawbacks when it comes to textual queries [106] (even speech recognition is challenging due to the noise of other teams), while remote users with a normal and familiar keyboard (non-laptop), and a high-performance PC-setting, seem to be able to produce significantly more queries than mobile team members on site. However, `vitrivr-VR`, the sole VR system in the competition demonstrated promising potential for video browsing in VR. Despite often lacking the correct search item in the result set, it successfully located the correct item through browsing in nearly all KIS tasks, ultimately achieving a fifth-place ranking. This highlights the potential of VR user interfaces and underscores the importance of addressing their inherent

challenges, as this is an emerging and promising area of research that would enable the creation of more inclusive and versatile interactive video retrieval platforms.

Notably, teams sometimes failed to find the correct shot, although the right video (with another shot) was ranked in their top ten results. This clearly demonstrates the importance of user interface design. Future developments should focus not only on refining search engines but also on creating interfaces that empower users to inspect and filter diverse types of information efficiently.

VII. CONCLUSION

In this paper, we performed an extensive evaluation of the Video Browser Showdown 2023 (VBS2023), which took place in Bergen, Norway in January 2023 at the International Conference on MultiMedia Modeling (MMM2023). 13 teams from 10 different countries participated in this challenging large-scale video search competition addressing 7 AVS tasks and 19 KIS tasks. Our evaluation encompassed an examination of the participating systems, offering an overview of their methodologies and delineating both commonalities and distinctive features. Furthermore, we meticulously analyze system logs containing all user queries and results during the competition. This analysis offers a comparison of the systems' performance and characteristics from various perspectives, including submission speed, retrieval success, and employed query types. Moreover, it provided valuable insights into the strengths, challenges, and future research directions of modern video search. Overall, despite all the progress in semantic content understanding, performing specific content search tasks in large and diverse datasets remains challenging. The VBS provides a valuable platform to evaluate the true practical search performance and will continue to extend its tasks with different test tasks (e.g., question answering) and datasets (e.g., medical video data).

ACKNOWLEDGMENT

This manuscript reflects only the authors' views and opinions; neither the European Union nor the other Granting Authorities can be considered responsible for them.

REFERENCES

[1] *3D-ResNets-PyTorch*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/kenshohara/3D-ResNets-PyTorch
[2] *ALADIN*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/mesnico/ALADIN
[3] *BLIP*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/salesforce/BLIP
[4] *Cineast Retrieval and Feature Extraction Engine*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/vitrivr/cineast
[5] *CLIP*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/openai/CLIP
[6] *CLIP2Video*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/CryhanFang/CLIP2Video
[7] *CLIP4Clip*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/ArrowLuo/CLIP4Clip
[8] *Cottontail Database*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/vitrivr/cottontaildb
[9] *EfficientNetV2 Model*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/CSAILVision/places365
[10] *FasterRCNN+InceptionResNetV2 Network Trained on Open Images V4*. Accessed: Jan. 25, 2024. [Online]. Available: https://www.kaggle.com/models/google/faster-rcnn-inception-resnet-v2/frameworks/tensorFlow1/variations/faster-rcnn-openimages-v4-inception-resnet-v2/versions/1?tfhub-redirect=true
[11] *MMDetection*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/open-mmlab/mmdetection
[12] *Model Card for CLIP ViT-B/32—LAION-2B*. Accessed: Jan. 25, 2024. [Online]. Available: https://huggingface.co/laion/CLIP-ViT-B-32-laion2B-s34B-b79K
[13] *Model Card for CLIP ViT-B/32 Xlm Roberta Base—LAION-5B*. Accessed: Jan. 25, 2024. [Online]. Available: https://huggingface.co/laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k
[14] *Model Card for CLIP ViT-L/14—LAION-2B*. Accessed: Jan. 25, 2024. [Online]. Available: https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K
[15] *Model Card for Clip vit-Large Patch 14-336*. Accessed: Jan. 25, 2024. [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14-336
[16] *Multilingual Sentence & Image Embeddings With BERT*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/UKPLab/sentence-transformers
[17] *Official Implementation of Character Region Awareness for Text Detection*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/clovaai/CRAFT-pytorch
[18] *The Places365-CNNs for Scene Classification*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/CSAILVision/places365
[19] *Robust Speech Recognition via Large-Scale Weak Supervision*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/openai/whisper
[20] *TextToVideoRetrieval-TtimesV*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/bmezaris/TextToVideoRetrieval-TtimesV
[21] *Torchvision Models*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/pytorch/vision/tree/main/torchvision/models
[22] *TransNet V2: Shot Boundary Detection Neural Network*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/soCzech/TransNetV2
[23] *VISIONE*. Accessed: Mar. 28, 2024. [Online]. Available: https://github.com/aimh-lab/visione
[24] *Vitrivir's Custom Cross-Modal Network*. Accessed: Jan. 25, 2024. [Online]. Available: https://data.vitrivr.org/VisualTextCoEmbedding/inception_resnet_v2_weights_tf_dim_ordering_tf_kernels_notop.tar.gz
[25] *VitriVR Interface*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/vitrivr/vitrivr-ng
[26] *Vitrivr-VR Interface*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/vitrivr/vitrivr-vr
[27] *YOLOv4—Neural Networks for Object Detection (Windows and Linux Version of Darknet)*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/kiyoshiiriemon/yolov4_darknet
[28] *YOLOv5*. Accessed: Jan. 25, 2024. [Online]. Available: https://github.com/ultralytics/yolov5
[29] L. Agnolucci, A. Baldrati, M. Bertini, and A. Del Bimbo, "Zero-shot image retrieval with human feedback," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 9417–9419.
[30] G. Amato, P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, and C. Vairo, "The VISIONE video search system: Exploiting off-the-shelf text search engines for large-scale video retrieval," *J. Imag.*, vol. 7, no. 5, p. 76, Apr. 2021.
[31] G. Amato, P. Bolettieri, F. Carrara, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, and C. Vairo, "VISIONE at video browser showdown 2023," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2023, pp. 615–621.
[32] G. Amato, F. Carrara, F. Falchi, C. Gennaro, and L. Vadicamo, "Large-scale instance-level image retrieval," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102100.
[33] S. Andreadis, A. Moumtzidou, D. Galanopoulos, N. Pantelidis, K. Apostolidis, D. Touska, K. Gkountakos, M. Pegia, I. Gialampoukidis, S. Vrochidis, V. Mezaris, and I. Kompatsiaris, "VERGE in VBS 2022," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2022, pp. 530–536.
[34] R. Arnold, L. Sauter, and H. Schuldt, "Free-form multi-modal multimedia retrieval (4MR)," in *Proc. 29th Int. Conf. MultiMedia Modeling (MMM)*, Bergen, Norway. Berlin, Germany: Springer, Jan. 2023, pp. 678–683.

[35] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9357–9366.

[36] W. Bailer, R. Arnold, V. Benz, D. Coccomini, A. Gkagkas, G. Guomundsson, S. Heller, B. Jónsson, J. Lokoc, N. Messina, N. Pantelidis, and J. Wu, "Improving query and assessment quality in text-based interactive video retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Retr.* New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 597–601.

[37] K. U. Barthel, N. Hezel, K. Jung, and K. Schall, "Improved evaluation and generation of grid layouts using distance preservation quality and linear assignment sorting," *Comput. Graph. Forum*, vol. 42, no. 1, pp. 261–276, Feb. 2023.

[38] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, "Food-101—Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 446–461.

[39] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, Sep. 2018, pp. 833–851.

[41] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7270–7292, Jun. 2023.

[42] Z. Chen and B. Zhu, "Some formal analysis of Rocchio's similarity-based relevance feedback algorithm," *Inf. Retr.*, vol. 5, pp. 61–86, Jan. 2002.

[43] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos, "PicHunter: Bayesian relevance feedback for image retrieval," in *Proc. 13th Int. Conf. Pattern Recognit.*, vol. 3, Aug. 1996, pp. 361–369.

[44] M. Crucianu, M. Ferecatu, and N. Boujemaa, "Relevance feedback for image retrieval: A short survey," in *Proc. Rep. DELOS2 Eur. Netw. Excellence (FP6)*, 2004, pp. 17–24.

[45] H. Fang, P. Xiong, L. Xu, and Y. Chen, "CLIP2Video: Mastering video-text retrieval via image CLIP," 2021, *arXiv:2106.11097*.

[46] D. Galanopoulos and V. Mezaris, "Are all combinations equal? Combining textual and visual features with multiple space learning for text-based video retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2023, pp. 627–643.

[47] R. Gasser, L. Rossetto, and H. Schuldt, "Multimodal multimedia retrieval with vitrivr," in *Proc. Int. Conf. Multimedia Retr.*, Ottawa, ON, Canada, Jun. 2019, pp. 391–394.

[48] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[49] K. Gkountakos, D. Touska, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Spatio-temporal activity detection and recognition in untrimmed surveillance videos," in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 451–455.

[50] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[51] C. Gurrin, B. Jónsson, D. T. D. Nguyen, G. Healy, J. Lokoc, L. Zhou, L. Rossetto, M.-T. Tran, W. Hürst, W. Bailer, and K. Schoeffmann, "Introduction to the sixth annual lifelog search challenge, LSC'23," in *Proc. 2023 ACM Int. Conf. Multimedia Retr.* New York, NY, USA: Association for Computing Machinery, 2023, pp. 678–679.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] X. He, O. King, W.-Y. Ma, M. Li, and H.-J. Zhang, "Learning a semantic space from user's relevance feedback for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 39–48, Jan. 2003.

[55] S. Heller, R. Arnold, R. Gasser, V. Gsteiger, M. Parian-Scherb, L. Rossetto, L. Sauter, F. Spiess, and H. Schuldt, "Multi-modal interactive video retrieval with temporal queries," in *Proc. 28th Int. Conf. MultiMedia Model.*, Phu Quoc, Vietnam, in Lecture Notes in Computer Science, vol. 13142. Cham, Switzerland: Springer, 2022, pp. 493–498.

[56] S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. Jónsson, J. Lokoč, A. Leibetseder, F. Mejzlík, L. Peška, L. Rossetto, K. Schall, K. Schoeffmann, H. Schuldt, F. Spiess, L.-D. Tran, L. Vadicamo, P. Veselý, S. Vrochidis, and J. Wu, "Interactive video retrieval evaluation at a distance: Comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 1, pp. 1–18, Mar. 2022.

[57] N. Hezel, K. U. Barthel, K. Schall, and K. Jung, "Fast approximate nearest neighbor search with a dynamic exploration graph using continuous refinement," 2023, *arXiv:2307.10479*.

[58] N. Hezel, K. Schall, K. Jung, and K. U. Barthel, "Efficient search and browsing of large-scale video collections with vibro," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2022, pp. 487–492.

[59] N. Hoang-Xuan, E.-R. Nguyen, T.-L. Nguyen-Ho, M.-K. Pham, Q.-T. Nguyen, H.-P. Trang-Trung, V.-T. Ninh, Tu-Khiem Le, C. Gurrin, and M.-T. Tran, "V-FIRST 2.0: Video event retrieval with flexible textual-visual intermediary for VBS 2023," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2023, pp. 652–657.

[60] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "OpenCLIP," Zenodo, Version 0.1, Jul. 2021. [Online]. Available: https://github.com/mlfoundations/open_clip/tree/v2.0.2, doi: 10.5281/zenodo.5143773.

[61] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[62] G. Jocher et al., 2022, "Ultralytics/yolov5: v6.2—YOLOv5 classification models, Apple M1, Reproducibility, ClearML and Deci.AI integrations, *Zenodo*, doi: 10.5281/zenodo.7002879.

[63] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[64] O. S. Khan, B. Jónsson, J. Zahálka, S. Rudinac, and M. Worring, "Impact of interaction strategies on user relevance feedback," in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 590–598.

[65] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[66] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[67] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong, "W2VV++: Fully deep learning for ad-hoc video search," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1786–1794.

[68] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, "TGIF: A new dataset and benchmark on animated GIF description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4641–4650.

[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[70] J. Lokoč, S. Andreadis, W. Bailer, A. Duane, C. Gurrin, Z. Ma, N. Messina, T.-N. Nguyen, L. Peška, L. Rossetto, L. Sauter, K. Schall, K. Schoeffmann, O. S. Khan, F. Spiess, L. Vadicamo, and S. Vrochidis, "Interactive video retrieval in the age of effective joint embedding deep models: Lessons from the 11th VBS," *Multimedia Syst.*, vol. 29, pp. 3481–3504, Aug. 2023.

[71] J. Lokoč, Z. Vopálková, P. Dokoupil, and L. Peška, "Video search with CLIP and interactive text query reformulation," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2023, pp. 628–633.

[72] J. Lokoč, W. Bailer, K. U. Barthel, C. Gurrin, S. Heller, B. Jónsson, L. Peška, L. Rossetto, K. Schoeffmann, L. Vadicamo, S. Vrochidis, and J. Wu, "A task category space for user-centric comparative multimedia search evaluations," in *Proc. Int. Conf. Multimedia Model.*, 2022, pp. 193–204.

[73] J. Lokoč, G. Kovalč ik, T. Souč ek, J. Moravec, and P. Čech, "A framework for effective known-item search in video," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1777–1785.

[74] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad, "On influential trends in interactive video retrieval: Video browser showdown 2015–2017," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3361–3376, Dec. 2018.

[75] J. Lokoc and L. Peska, "A study of a cross-modal interactive search tool using clip and temporal fusion," in *Proc. 29th Int. Conf. MultiMedia Model.*, Bergen, Norway, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, Jan. 2023.
[76] S. Lubos, M. Rubino, C. Tautschnig, M. Tautschnig, B. Wen, K. Schoeffmann, and A. Felfernig, "Perfect match in video retrieval," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2023, pp. 634–639.
[77] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, Oct. 2022.
[78] Z. Ma, J. Wu, Z. Hou, and C.-W. Ngo, "Reinforcement learning-based interactive video search," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2022, pp. 549–555.
[79] Z. Ma, J. Wu, W. Loo, and C.-W. Ngo, "Reinforcement learning enhanced PicHunter for interactive search," in *Proc. Conf. Multimedia Model.*, 2023, pp. 690–696.
[80] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 181–196.
[81] N. Messina, M. Stefanini, M. Cornia, L. Baraldi, F. Falchi, G. Amato, and R. Cucchiara, "ALADIN: Distilling fine-grained alignment scores for efficient image-text matching and retrieval," in *Proc. 19th Int. Conf. Content-Based Multimedia Indexing*, Sep. 2022, pp. 64–70.
[82] T.-N. Nguyen, B. Puangthamawathanakun, A. Caputo, G. Healy, B. T. Nguyen, C. Arpnikanondt, and C. Gurrin, "VideoCLIP: An interactive CLIP-based video retrieval system at VBS2023," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2023, pp. 671–677.
[83] T.-N. Nguyen, B. Puangthamawathanakun, G. Healy, B. T. Nguyen, C. Gurrin, and A. Caputo, "Videofall—A hierarchical search engine for VBS2022," in *Proc. 28th Int. Conf. MultiMedia Modeling*, Phu Quoc, Vietnam. Cham, Switzerland: Springer, 2022, pp. 518–523.
[84] N. Pantelidis, S. Andreadis, M. Pegia, A. Moumtzidou, D. Galanopoulos, K. Apostolidis, D. Touska, K. Gkountakos, I. Gialampoukidis, S. Vrochidis, V. Mezaris, and I. Kompatsiaris, "VERGE in VBS 2023," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2023, pp. 658–664.
[85] N. Pittaras, F. Markatopoulou, V. Mezaris, and I. Patras, "Comparison of fine-tuning and extension strategies for deep convolutional neural networks," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2017, pp. 102–114.
[86] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 8748–8763.
[87] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 202, Jul. 2023, pp. 28492–28518.
[88] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5106–5115.
[89] L. Rossetto, R. Gasser, J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, T. Soucek, P. A. Nguyen, P. Bolettieri, A. Leibetseder, and S. Vrochidis, "Interactive video retrieval in the age of deep learning—Detailed evaluation of VBS 2019," *IEEE Trans. Multimedia*, vol. 23, pp. 243–256, 2021.
[90] L. Rossetto, R. Gasser, L. Sauter, A. Bernstein, and H. Schuldt, "A system for interactive multimedia retrieval evaluations," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2021, pp. 385–390.
[91] L. Rossetto, R. Gasser, and H. Schuldt, "Query by semantic sketch," 2019, *arXiv:1909.12526*.
[92] L. Rossetto, M. A. Parian, R. Gasser, I. Giangreco, S. Heller, and H. Schuldt, "Deep learning-based concept detection in vitrivr," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2019, pp. 616–621.
[93] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt, "V3C—A research video collection," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2019, pp. 349–360.
[94] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[95] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," *Knowl. Eng. Rev.*, vol. 18, no. 2, pp. 95–145, Jun. 2003.
[96] L. Sauter, R. Gasser, S. Heller, L. Rossetto, C. Saladin, F. Spiess, and H. Schuldt, "Exploring effective interactive text-based video search in vitrivr," in *Proc. 29th Int. Conf. MultiMedia Model.*, Bergen, Norway, in Lecture Notes in Computer Science, vol. 13833. Cham, Switzerland: Springer, Jan. 2023, pp. 646–651.
[97] K. Schall, K. U. Barthel, N. Hezel, and. Jung, "GPR1200: A benchmark for general-purpose content-based image retrieval," in *Proc. 28th Int. Conf., MultiMedia Modeling*, Phu Quoc, Vietnam. Berlin, Germany: Springer, Jun. 2022, pp. 205–216.
[98] K. Schall, K. U. Barthel, N. Hezel, and K. Jung, "Improving image encoders for general-purpose nearest neighbor search and classification," in *Proc. ACM Int. Conf. Multimedia Retr.* New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 57–66.
[99] K. Schall, N. Hezel, K. Jung, and K. U. Barthel, "Vibro: Video browsing with semantic and visual image embeddings," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2023, pp. 665–670.
[100] K. Schoeffmann, D. Stefanics, and A. Leibetseder, "DiveXplore at the video browser showdown 2023," in *Proc. 29th Int. Conf. MultiMedia Modeling*, Bergen, Norway. Cham, Switzerland: Springer, Jan. 2023, pp. 684–689.
[101] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 25278–25294.
[102] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.
[103] W. Song, J. He, X. Li, S. Feng, and C. Liang, "QIVISE: A quantum-inspired interactive video search engine in VBS2023," in *Proc. 29th Int. Conf. MultiMedia Modeling*, Bergen, Norway. Cham, Switzerland: Springer, Jan. 2023, pp. 640–645.
[104] T. Souč ek and J. Lokoč, "TransNet v2: An effective deep network architecture for fast shot transition detection," 2020, *arXiv:2008.04838*.
[105] F. Spiess, R. Gasser, S. Heller, M. Parian-Scherb, L. Rossetto, L. Sauter, and H. Schuldt, "Multi-modal video retrieval in virtual reality with vitrivr-VR," in *Proc. Int. Conf. Multimedia Model.*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2022, pp. 499–504.
[106] F. Spiess, R. Gasser, S. Heller, H. Schuldt, and L. Rossetto, "A comparison of video browsing performance between desktop and virtual reality interfaces," in *Proc. ACM Int. Conf. Multimedia Retr.*, Thessaloniki, Greece, Jun. 2023, pp. 535–539.
[107] F. Spiess, S. Heller, L. Rossetto, L. Sauter, P. Weber, and H. Schuldt, "Traceable asynchronous workflows in video retrieval with vitrivr-VR," in *Proc. 29th Int. Conf. MultiMedia Model.*, in Lecture Notes in Computer Science, vol. 13833, Bergen, Norway. Cham, Switzerland: Springer, 2023, pp. 622–627.
[108] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[109] L.-D. Tran, M.-D. Nguyen, D.-T. Dang-Nguyen, S. Heller, F. Spiess, J. Lokoc, L. Peška, T.-N. Nguyen, O. S. Khan, A. Duane, B. Jónsson, L. Rossetto, A.-Z. Yen, A. Alateeq, N. Alam, M.-T. Tran, G. Healy, K. Schoeffmann, and C. Gurrin, "Comparing interactive retrieval approaches at the lifelog search challenge 2021," *IEEE Access*, vol. 11, pp. 30982–30995, 2023.
[110] M.-T. Tran, N. Hoang-Xuan, H.-P. Trang-Trung, T.-C. Le, M.-K. Tran, M.-Q. Le, T.-K. Le, V.-T. Ninh, and C. Gurrin, "V-FIRST: A flexible interactive retrieval system for video at VBS 2022," in *Proc. 28th Int. Conf. MultiMedia Modeling*, Phu Quoc, Vietnam. Cham, Switzerland: Springer, Jun. 2022, pp. 562–568.
[111] Q.-T. Truong, T.-A. Vu, T.-S. Ha, J. Lokoc, Y. H. W. Tim, A. Joneja, and S.-K. Yeung, "Marine video kit: A new marine video dataset for content-based analysis and retrieval," in *Proc. 29th Int. Conf. MultiMedia Model.*, Bergen, Norway. Cham, Switzerland: Springer, Jan. 2023, pp. 539–550.
[112] S. Uprety, D. Gkoumas, and D. Song, "A survey of quantum theory inspired approaches to information retrieval," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–39, Sep. 2021.
[113] J. Wang et al., "Milvus: A purpose-built vector data management system," in *Proc. Int. Conf. Manage. Data*, Jun. 2021, pp. 2614–2627.

[114] P. Wang, Y. Hou, Z. Li, and Y. Zhang, "QIRM: A quantum interactive retrieval model for session search," *Neurocomputing*, vol. 451, pp. 57–66, Sep. 2021.
[115] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4580–4590.
[116] J. Wu and C.-W. Ngo, "Interpretable embedding for ad-hoc video search," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2020, pp. 3357–3366.
[117] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.
[118] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.
[119] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, "VarifocalNet: An IoU-aware dense object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8510–8519.
[120] R. Raturi, "Adapting deep features for scene recognition utilizing places database," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 487–495.

LUCIA VADICAMO received the master's degree in mathematics and the Ph.D. degree in information engineering from the University of Pisa, in 2013 and 2018, respectively. She is currently a Researcher with the Information Science and Technologies Institute (ISTI), National Research Council (CNR), Pisa, Italy. She is also leading a team dedicated to the development of the VISIONE System. Her research interests include multimedia information retrieval, artificial intelligence, and similarity search.

RAHEL ARNOLD received the master's degree, in 2022. She is currently pursuing the Ph.D. degree with the Department of Mathematics and Computer Science, University of Basel. Her current research interests include multimedia retrieval, mixed reality, and a combination thereof.

WERNER BAILER (Member, IEEE) received the Dipl.-Ing. degree in media technology and design, in 2002, with a diploma thesis on motion estimation and segmentation. He is currently a Key Researcher with the DIGITAL–Institute for Digital Technologies, Joanneum Research, Graz, Austria. His research interests include audiovisual content analysis and retrieval, media production technologies, and machine learning, contributing also to standardization in these areas.

FABIO CARRARA received the master's and Ph.D. degrees in computer engineering from the University of Pisa, Italy, in 2015 and 2019, respectively. He is currently a Researcher with the Information Science and Technologies Institute (ISTI), National Research Council (CNR), Pisa, Italy. His research interests include deep learning for multimedia data with a focus on visual perception, image classification, and content-based and cross-media image retrieval and analysis.

CATHAL GURRIN is currently a Professor with the School of Computing, Dublin City University (DCU), and the Head of the Adapt Centre at DCU. He is the Founder of the Annual ACM Lifelog Search Challenge and a Co-Organizer of the Annual Video Browser Showdown. He is interested in building rich multimodal user models and deploying them to solve real-world challenges using AI. His research interests include personal media analytics, user modeling, and lifelogging.

NICO HEZEL received the B.Sc. and M.Sc. degrees in international media informatics, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree in graph-based nearest neighbor search. Afterwards, he started teaching machine learning and several visual computing-related courses at the HTW Berlin–University of Applied Sciences. He is also a Researcher with the Visual-Computing Laboratory, HTW Berlin–University of Applied Sciences.

XINGHAN LI is currently pursuing the bachelor's degree with the School of Computer Science, Wuhan University. He serves as the Core Developer of the QIVISE system. His research interests include video retrieval, including the application of quantum information theory in multimodal retrieval and interactive multimedia.

JAKUB LOKOC is currently an Associate Professor with Charles University, Prague, Czech Republic. He is a Co-Organizer of the Video Browser Showdown and Lifelog Search Challenges. So far, he has coauthored more than 100 peer-reviewed articles. His research interests include similarity searching, multimedia indexing, and interactive video retrieval.

SEBASTIAN LUBOS (Graduate Student Member, IEEE) received the master's degree, in 2021. He is currently pursuing the Ph.D. degree with the Applied Software Engineering and AI Research Group, Graz University of Technology, Austria. His current research interest includes the effective and efficient retrieval and recommendation of videos, with a special focus on educational videos.

ZHIXIN MA is currently pursuing the Ph.D. degree with the School of Computing and Information Systems, Singapore Management University. His general research interest includes multimedia computing and analysis, with a specific focus on interactive multimedia search and user behavior simulation. He is a Core Member of the VIREO Team.

NICOLA MESSINA received the master's degree in computer engineering and the Ph.D. degree in information engineering from the University of Pisa, in 2018 and 2022, respectively. He is currently a Researcher with the Information Science and Technologies Institute (ISTI), National Research Council (CNR), Pisa, Italy. He is researching deep learning methods for relational understanding in multimedia data, with particular emphasis on transformer-based architectures for effective and efficient cross-modal analysis and retrieval of images, texts, and videos.

THAO-NHU NGUYEN is currently pursuing the Ph.D. degree in computer science with the School of Computing, Dublin City University. She is the Core Developer of the videoclip system. Her research interests include multimedia analysis, representation, and retrieval, especially for lifelog data.

LADISLAV PESKA is currently an Assistant Professor with Charles University, Prague, Czech Republic. He has coauthored more than 80 peer-reviewed articles. His research interests include recommender systems, personalized information retrieval, and interactive video retrieval. He is a member of the CVHunter Team.

LUCA ROSSETTO received the Ph.D. degree in computer science from the University of Basel, in 2018. He is currently a Postdoctoral Researcher with the Department of Informatics, University of Zurich. He is a Core Contributor to the Vitrivr Project as well as the DRES retrieval evaluation system. His research interests include the analysis, management, and retrieval of multimedia data.

LORIS SAUTER received the Ph.D. degree in computer science from the University of Basel, in 2024. He is currently a Postdoctoral Researcher with the Department of Mathematics and Computer Science, University of Basel. He is a Core Contributor to the DRES retrieval evaluation system and the Vitrivr Project. He is actively contributing to other research projects at the Databases and Information Systems Research Group. His research interests include multimedia retrieval, multimedia retrieval evaluations, and the analysis thereof.

KLAUS SCHÖFFMANN (Member, IEEE) is currently an Associate Professor with the Institute of Information Technology (ITEC), Universität Klagenfurt, Austria. He has coauthored more than 100 publications on various topics in multimedia, inclusive of many works on different aspects of medical video analysis. His research interests include video content understanding (in particular medical/surgery videos), multimedia retrieval, interactive multimedia, and applied deep learning. He is a member of the ACM and a regular reviewer for international conferences and journals in the field of multimedia.

FLORIAN SPIESS is currently pursuing the Ph.D. degree with the Databases and Information Systems Research Group, Department Mathematics and Computer Science, University of Basel. He is the Lead Developer of the vitrivr-VR virtual reality multimedia retrieval system and a Core Contributor to the Vitrivr Project. His research interest includes multimedia analytics using virtual reality interfaces.

MINH-TRIET TRAN received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Science–VNUHCM, in 2001, 2005, and 2009, respectively. In 2001, he joined the University of Science. He was a Visiting Scholar with the National Institutes of Informatics (NII), Japan, from 2008 to 2010, and the University of Illinois at Urbana–Champaign (UIUC), from 2015 to 2016. He is currently the Vice President with the University of Science–VNUHCM. His research interests include cryptography, security, computer vision, and human–computer interaction. He is a Membership Development and Student Activities Coordinator of the IEEE Vietnam. He is also a member of the Advisory Council for Artificial Intelligence Development of Ho Chi Minh City and the Vice Chairperson of Vietnam Information Security Association (VNISA, South Branch).

STEFANOS VROCHIDIS (Member, IEEE) received the Diploma degree in electrical engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2000, the M.Sc. degree in radio frequency communication systems from the University of Southampton, Southampton, U.K., in 2001, and the Ph.D. degree in electronic engineering from the Queen Mary University of London, London, U.K., in 2013. He is currently a Senior Researcher (Grade C) with the Multimedia Knowledge and Social Media Analytics Laboratory, Centre for Research and Technology Hellas, Information Technologies Institute, Thermi-Thessaloniki, Greece, and the Head of the Multimodal Data Fusion and Analytics Group. His research interests include multimedia analysis and retrieval, multimodal fusion, computer vision, multimodal analytics based on artificial intelligence, semantic web, as well as media and arts, environmental, and security applications.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE-ITALY' within the CRUI CARE Agreement