WISE: A Multimodal Search Engine for Visual Scenes, Audio, Objects, Faces, Speech, and Metadata

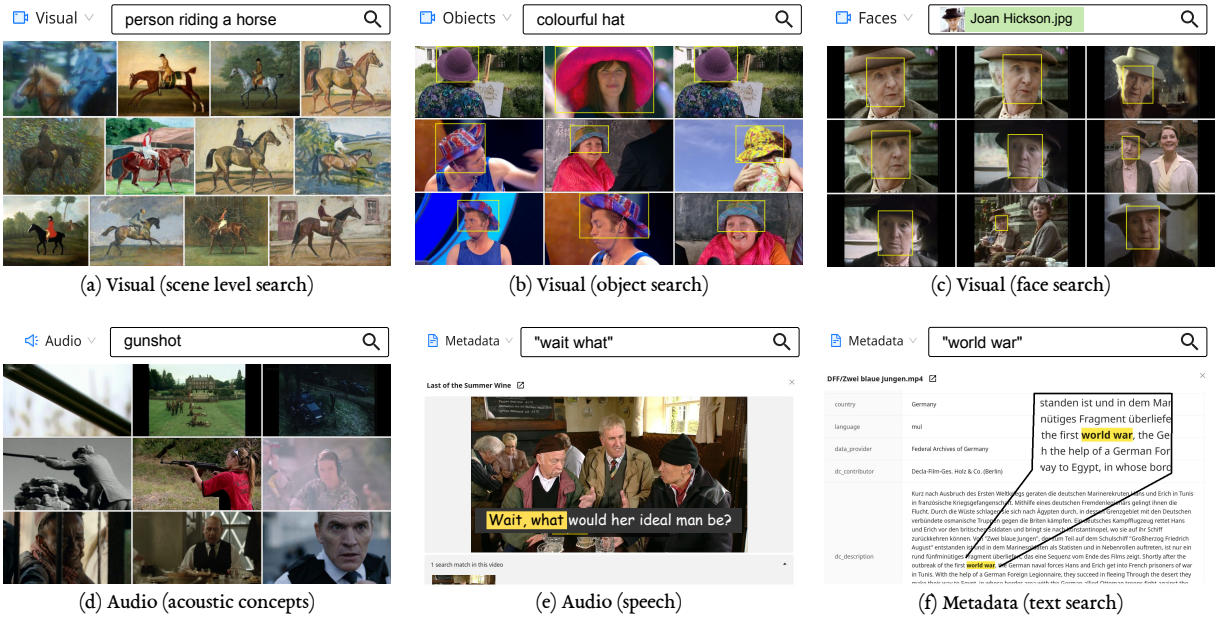
Prasanna Sridhar
prasanna@robots.ox.ac.uk
Engineering Science
University of Oxford
Oxford, UK

Horace Lee
horacelee@robots.ox.ac.uk
Engineering Science
University of Oxford
Oxford, UK

David M. S. Pinto
pinto@robots.ox.ac.uk
Engineering Science
University of Oxford
Oxford, UK

Andrew Zisserman
az@robots.ox.ac.uk
Engineering Science
University of Oxford
Oxford, UK

Abhishek Dutta
adutta@robots.ox.ac.uk
Engineering Science
University of Oxford
Oxford, UK



(a) Visual (scene level search) (b) Visual (object search) (c) Visual (face search)

(d) Audio (acoustic concepts) (e) Audio (speech) (f) Metadata (text search)

Figure 1: WISE enables search across visual, audio, and metadata streams. Visual search operates at both the scene level (*e.g.* person riding a horse) and the object level (*e.g.* colourful hat) with support for finding specific faces (*e.g.* using a photo of an actress). Audio search retrieves audiovisual segments based solely on acoustic content (*e.g.* gunshot) or speech (*e.g.* "wait what"). WISE also supports traditional text-based search over media-level metadata.

Abstract

In this paper, we present WISE, an open-source audiovisual search engine which integrates a range of multimodal retrieval capabilities into a single practical tool, accessible to users without machine learning expertise. WISE supports natural-language and reverse-image queries at both the scene level (*e.g.* empty street) and object level (*e.g.* horse) across images and videos; face-based search for specific individuals; audio retrieval of acoustic events using text (*e.g.* wood creak) or an audio file; search over automatically transcribed speech; and filtering by user-provided metadata. Rich insights can be obtained by combining queries across modalities — for example, retrieving German trains from a historical archive by applying the object query "train" and the metadata query "Germany", or searching for a face in a place. By employing vector search techniques, WISE can scale to support efficient retrieval over millions of images or thousands of hours of video. Its modular architecture facilitates the integration of new audio or visual models. WISE can be deployed locally for private or sensitive collections, and has been applied to a number of disparate real-world use cases. Code is available at https://gitlab.com/vgg/wise/wise.

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808375

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; **Image search**; **Video search**; **Speech / audio search**; **Search interfaces**.

Keywords

Multimodal retrieval, Audiovisual Search, Video Search, Audio Search, Object Search, Face Search, Speech Search

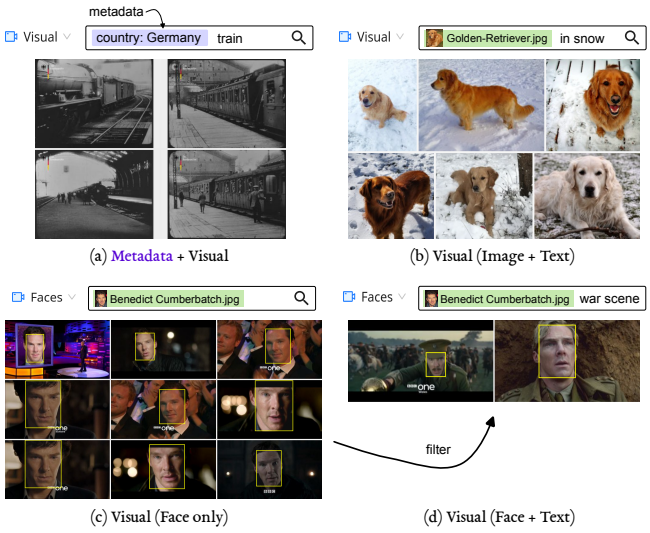
ACM Reference Format:
Prasanna Sridhar, Horace Lee, David M. S. Pinto, Andrew Zisserman, and Abhishek Dutta. 2026. WISE: A Multimodal Search Engine for Visual Scenes, Audio, Objects, Faces, Speech, and Metadata. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805712.3808375

1 Introduction

Large-scale audiovisual collections are now everywhere, powered by a new wave of devices (*e.g.* mobile phones) and online platforms (*e.g.* social networks) that make it easy to capture, store and share rich multimedia content — images, audio and video - on a massive scale. Although automatically generated metadata (*e.g.* location, date) and manually annotated metadata (*e.g.* title, description, tags) help users explore these collections, they are often limited and cannot fully describe the content. Moreover, collecting high-quality metadata becomes expensive and difficult at scale. In order to be useful, such large collections need search tools that can go beyond metadata and leverage the underlying visual and audio content. While recent advances in multimodal models offer powerful capabilities for content discovery, they remain largely inaccessible to non-technical users due to the expertise required for deployment and integration. This paper introduces WISE (WISE Search Engine), a user-friendly tool which supports searching across visual, audio, and metadata information streams of an audiovisual collection. WISE also offers composite multimodal search capability (*e.g.* metadata + visual or face + visual) as described in Section 2. The workflow adopted by WISE for processing audiovisual data is defined in Section 3 and the software architecture is detailed in Section 4. The case studies described in Section 5 show the impact of the WISE open source software on various research disciplines and industrial sectors. Audiovisual search engines are essential for curating, searching and managing large collections of multimedia content that have become generally available in most research and commercial avenues. The WISE software introduced in this paper has the potential to become a platform for all general purpose audiovisual search requirements.

2 Multimodal Search Capabilities

Visual Search. The visual stream consists of images and frames extracted from videos. WISE supports visual search at two levels. The first is scene level search, which uses queries that describe the overall composition of a scene. For example, the natural language query "person riding a horse" applied to a collection of paintings retrieves images depicting this scene, as shown in Figure 1a. An image can also be used as a search query to find visually similar scenes. The second is object-level search, which provides a more



(a) Metadata + Visual (b) Visual (Image + Text)

(c) Visual (Face only) (d) Visual (Face + Text)

Figure 2: WISE supports multimodal search which allows (a) filtering visual search results using metadata, or (b) performing composed queries by combining an image with a refining text description. Similarly, a compositional query can be applied to face search results (c) to retrieve (d) an actor in a particular type of scene, for example.

fine grained view of the objects within a scene. For example, the query "colourful hat" retrieves video segments or images containing hats of various colours and highlights the matched regions with bounding boxes as shown in Figure 1b. WISE also supports face search, which can be viewed as a special class of object search, where the results are retrieved by matching the unique identity in facial imagery. Figure 1c shows an example using an image of an actor's face as the query.

Audio Search. The audio stream in video or audio files can also be searched in WISE using keywords that represent acoustic concepts such as gunshot, laughter, siren, or footsteps. For example, the query "gunshot" retrieves video segments or audio files containing the sound of a gunshot, as shown in Figure 1d. Although some retrieved segments may also display visual depictions of guns, the results are based solely on the audio content. An audio file can also be used as a search query for audio search. The audio stream frequently includes human speech, and WISE enables search over spoken words using text queries. WISE uses Automatic Speech Recognition (ASR) to transcribe speech into metadata, allowing queries such as "wait what" to retrieve relevant video segments where those words are spoken, as illustrated in Figure 1e.

Metadata Search. Certain metadata (*e.g.* date, location) are typically added to media files automatically while other descriptive metadata (*e.g.* image captions and tags) often need to be added manually by human annotators. WISE supports full text search over this media-level metadata like traditional text-based search tools. For example, the query "world war" retrieves all media files whose metadata contains those terms, as shown in Figure 1f.

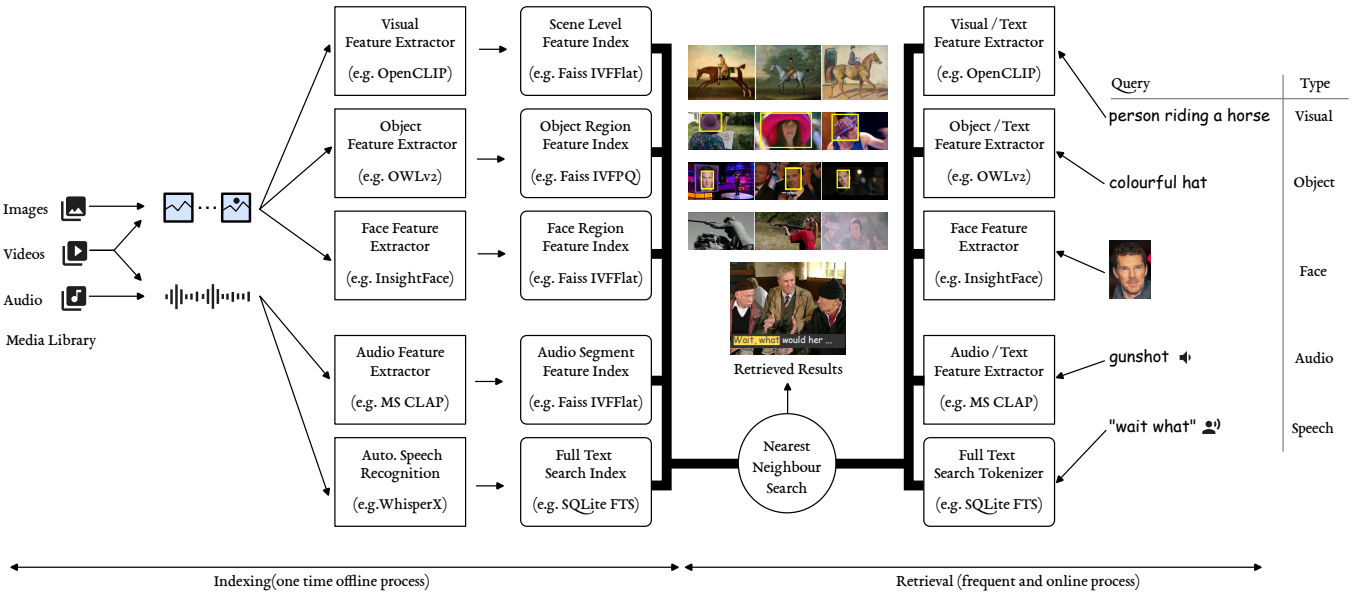


Figure 3: WISE organises the media library into two streams: visual and audio. Scene and region level features (or embeddings) are extracted from the visual stream, which includes images and video frames, while acoustic event and speech features are extracted from the audio stream. All features are saved in a vector store to enable fast nearest neighbour search. Feature extraction and indexing are performed once as an offline process. At query time, audiovisual features are extracted from the search query and matched against the index to retrieve semantically similar content from the media library.

Composite Search. WISE supports composite multimodal search by allowing users to combine search queries across modalities. For example, the query in Figure 2a applies the metadata filter "`country:Germany`" to a visual search with the natural language query "train". This retrieves video segments that contain a train and are tagged with "Germany" in the `country` metadata field. In many cases, queries are most naturally defined using both an image and a refining text description. This approach, known as composed retrieval [7], is illustrated in Figure 2b, where an image of a dog combined with the text "in snow" retrieves images of dogs in snowy scenes. Face search results, as shown in Figure 2c, can also be combined with visual search, enabling queries such as finding specific actors in a particular scene or location, as shown in Figure 2d.

3 Audiovisual Data Processing and Search

As shown in Figure 3 (left), WISE aggregates a media library into two streams: visual and audio. The visual stream consists of images and video frames sampled, for example, at 2 frames per second. The audio stream consists of audio snippets extracted from audio files and audio channel of videos using a window based sampling strategy (*e.g.* 4 s window with 2 s overlap). A set of visual feature extractors operate on the visual stream to extract features that can support content-based search at the scene level and region level (*e.g.* object, face). For example, OpenCLIP [9, 13] features are extracted from each sampled frame to support scene level visual search. Spatial region features are extracted by OWLv2 [12] and InsightFace [3] feature extractors to enable visual search based on objects and faces respectively. The scene level features provide a high level overview while the spatial region features offer a more detailed view of the visual stream. Similarly, audio feature extractors operate on the audio stream to extract features that support search for acoustic concepts and speech. For acoustic concepts (*e.g.* gunshot, footsteps, clapping), CLAP [5] features are extracted from each window. ASR models (*e.g.* WhisperX [1]) are applied to transcribe speech, enabling text-based search of spoken words.

The extracted features (*e.g.* 768 dimensional vectors) are stored in a vector search index (*e.g.* Faiss IndexIVFFlat [4]) that supports fast approximate nearest neighbour retrieval. For feature extractors that generate a large number of candidate regions (*e.g.* OWLv2), a more aggressive compression scheme (*e.g.* Faiss IndexIVFPQ) can be used to reduce compute and storage costs with a minor impact on retrieval accuracy. The relationship between each feature vector and its associated media content (*e.g.* filename, timestamp, region coordinates) are stored in a SQLite database. Metadata associated with each media file (*e.g.* caption, title) is indexed using traditional full text search (*e.g.* FTS search in SQLite [6]). Feature extraction and indexing are compute and storage intensive processes, but they only need to be performed once and can be executed offline.

The audiovisual feature extractors used by WISE map text, images, video frames, and audio into a shared vector space. In this space, audiovisual exemplars and their corresponding text descriptions lie close together, while unrelated descriptions are farther apart. These feature extractors are also called vision-language models or audio-language models, and they are trained on a large corpus of audiovisual exemplars and their corresponding text descriptions sourced from the internet. As shown in Figure 3 (right), search queries (image or text) are transformed into feature vectors using

the same set of feature extractors that were used during indexing. An approximate nearest neighbour algorithm retrieves indexed feature vectors that lie closest to the query vector thereby retrieving the temporal segment or spatial regions in the media library that are most semantically relevant. The retrieval step runs online and completes almost instantly, as it requires only one feature extraction and a fast nearest neighbour search.

4 Software Architecture

WISE follows a modular design consisting of five components: Loader, Extractor, Store, Index, and Search. The *Loader* reads media files and extracts audiovisual information in a format (*e.g.* tensor) that enables the *Extractor* to compute feature vectors which are persisted by the *Store*. The *Index* builds vector indexes from stored features to support fast nearest neighbour retrieval. This modular structure provides extensibility, scalability, and high performance.

Extensible. The Loader module can be extended to ingest new media formats (*e.g.* DICOM format from Ultrasound devices). New feature extractors can be integrated by implementing the Extractor's standard interface. For example, replacing the human face feature extractor with ChimpUFE [8] enables the creation of a search engine to study the behaviour and social network of chimpanzees [2]. The Store and Index can also be extended to use alternative storage and indexing backends. WISE's web frontend, built with React, supports straightforward extension of the user interface components.

Scalable. WISE has been deployed on collections as large as 55 million images from Wikimedia Commons [15] and over 6,000 hours of BBC videos. As shown in Figure 4, WISE can run in *aggregator mode*, distributing queries across multiple machines that each store a subset of the full media collection, and *merging* their results. This architecture allows WISE to scale millions of images and thousands of hours of video.

High-performance. The Extractor can run on a dedicated GPU machine (see Figure 4), handling batched extraction requests from multiple WISE instances to maximise GPU utilisation. The Loader uses multithreaded pre-fetching to deliver audiovisual data to the Extractor efficiently. A one hour video can be processed in under 10 minutes on a modern computer with a GPU. WISE returns results in under 1 second even for the large datasets aforementioned. Retrieval latency can be reduced further using more efficient indexes such as Product Quantisation with an Inverted File Index [11].

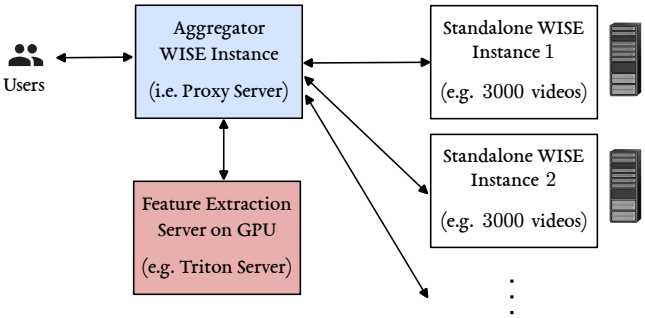


Figure 4: WISE can operate in aggregator mode, where a large audiovisual collection is split up across multiple standalone nodes each covering a different subset. A central instance distributes search queries to all standalone nodes, gathers their results, and presents a unified response to the user.

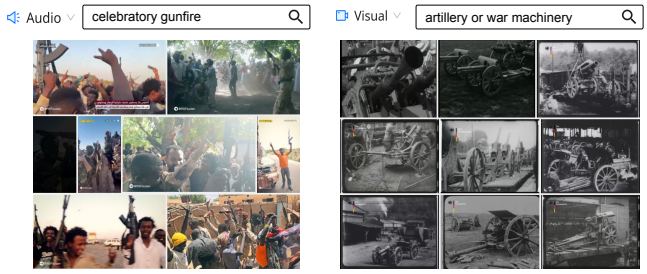
5 Case Studies

WISE has been adopted in many commercial and academic research projects. Public online demos are available[1] to showcase its deployment in different disciplines. Here are some examples.

Journalism. Open Source Intelligence (OSINT) is a growing trend in journalism which involves developing or investigating a story based on audiovisual content gathered from social media. WISE has been used as part of the reporting and documentary workflow of journalists dealing with a large audiovisual collection, as shown in Figure 5a.

Film and Media Research. WISE supports film and media scholars in curating and exploring historically significant audiovisual collections. An online demo developed for the Cinephile Challenge 2025 demonstrates its use on material from the Dutch and German national archives, as shown in Figure 5b.

Other Use Cases. Wildlife conservation organisations are using WISE to explore their video archives to identify relevant clips for environmental storytelling. Commercial archives employ WISE to improve retrieval and licensing of historically significant media.



(a) Investigative Journalism (b) Film and Media Research

Figure 5: WISE has proved invaluable for journalists investigating stories based on audiovisual content (left) and for film and media scholars studying historical archives (right).

6 Related Work

WISE is based on vision-language (*e.g.* [9, 13]) and audio-language (*e.g.* [5]) models that embed different modalities (*e.g.* visual, audio, text) in a shared embedding space, as described in Section 3. Some commercial entities (*e.g.* OpenAI, Cloud Providers) offer embedding model API services and leave the building of the search functionality and interfaces up to the user. Software-as-a-service multimedia search tools, *e.g.* Google Photos, TwelveLabs, and muse.ai, are more accessible to non-technical users, but often require a paid plan (when usage exceeds free tier limits) and rely on users to upload their data online. Other open source tools such as Exquisitor [14] and CineSearcher [10] also leverage vision-language models to

[1] https://www.robots.ox.ac.uk/~vgg/software/wise/examples/

enable natural language search over visual content, while delivering novel search functionalities like conversational search, interactive exploration and an iterative workflow for annotating multimedia content. In comparison, WISE offers search over additional specific visual content (faces, objects), audio and ASR; it enables search queries that combine modalities (*e.g.* a query composed of text and an image); and has a more flexible architecture allowing feature extractors to be easily changed as new and better models become available.

7 Discussion and Conclusion

WISE leverages recent advances in vision-language [9] and audio-language [5] models to enable exploration of audiovisual collections using text and image based queries. By allowing users to combine and filter search results from multiple modalities, our software has facilitated the use of audiovisual search in domains such as film and media research, and journalism which have so far relied mostly on manual curation and search. With a design that enables usage on a local machine, WISE allows users working with sensitive or proprietary collections to avoid the need to upload content to third party providers on the internet. The system can be extended to incorporate newly developed audiovisual models and search indexes. WISE is available open-source under the Apache License 2.0 and we hope that this work can serve as a practical tool across diverse audiovisual collections.

Acknowledgments

This work was funded by the EPSRC Programme Grant VisualAI EP/T028572/1. We are grateful to Dr. Ashish Thandavan for supporting the compute infrastructure requirements of this project, and the reviewers for their comments.

References

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech 2023*. 4489–4493. doi:10.21437/Interspeech.2023-78

[2] Max Bain, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owen, Kimberley J Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, et al. 2021. Automated audiovisual behavior recognition in wild primates. *Science advances* 7, 46 (2021), eabi4883.

[3] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.

[4] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[6] D. Richard Hipp. 2025. *SQLite*. https://sqlite.org/

[7] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. 2025. Collm: A large language model for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3994–4004.

[8] Vladimir Iashin, Horace Lee, Dan Schofield, and Andrew Zisserman. 2025. Self-supervised Learning on Camera Trap Footage Yields a Strong Universal Face Embedder. *arXiv preprint arXiv:2507.10552* (2025).

[9] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2025. *OpenCLIP*. doi:10.5281/zenodo.17171361

[10] Tobias Kreten and Marta Kipke. 2025. CineSearcher-A Multimodal Film Exploration Workspace. In *Proceedings of the 7th International Workshop on analySis, Understanding and proMotion of heritAge Contents*. 51–60.

[11] Yusuke Matsui, Yusuke Uchida, Hervé Jégou, and Shin'ichi Satoh. 2018. A survey of product quantization. *ITE Transactions on Media Technology and Applications* 6, 1 (2018), 2–10.

[12] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems* 36 (2023), 72983–73007.

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

[14] Ujjwal Sharma, Omar Shahbaz Khan, Stevan Rudinac, and Björn Þór Jónsson. 2025. Exquisitor at the Video Browser Showdown 2025: Unifying Conversational Search and User Relevance Feedback. In *MultiMedia Modeling*, Ichiro Ide, Ioannis Kompatsiaris, Changsheng Xu, Keiji Yanai, Wei-Ta Chu, Naoko Nitta, Michael Riegler, and Toshihiko Yamasaki (Eds.). Springer Nature Singapore, Singapore, 264–271.

[15] Prasanna Sridhar, Horace Lee, Abhishek Dutta, and Andrew Zisserman. 2023. WISE image search engine (WISE). In *Wiki workshop, virtual event*.