JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3.6607

PENCARIAN CERDAS ANTAR-MODA : EVOLUSI TEKNOLOGI *VIDEO-TEXT RETRIEVAL*

Muhammad Fiddiana Asyhari[*1], Fadli Dimas[2], Abib Maftuh Abu Bakar[3], Ade Bastian[4]
[1,2,3,4]Universitas Majalengka; Jl.K.H.Abdul Halim No.103,Majalengka,Jawa Barat,Indonesia;
Telp/Fax: (0233) 281496

Keywords:
Bibliometrik,
InternVid,
Text Retrieval,
Video-Text Retrieval
VOSviewer.

Corespondent Email:
muhamadfiddiana@gmail.com

Penelitian ini bertujuan untuk mengkaji tren ilmiah dan pendekatan mutakhir dalam bidang video-text retrieval melalui pendekatan analisis bibliometrik. Urgensi topik ini didorong oleh lonjakan pertumbuhan data audiovisual dan meningkatnya kebutuhan terhadap sistem pencarian yang mampu memahami keterkaitan semantik antara teks dan video. Data bibliometrik dikumpulkan menggunakan perangkat lunak Publish or Perish Harzing dengan sumber utama Google Scholar, kemudian divisualisasikan dan dianalisis menggunakan VOSviewer untuk mengidentifikasi klaster tematik, distribusi kata kunci, serta pola kolaborasi penulis. Dataset InternVid dijadikan sebagai referensi eksplorasi karena menyediakan jutaan klip video dengan anotasi teks yang kaya secara semantik. Hasil analisis menunjukkan lima klaster utama yang menggambarkan arah tematik dalam bidang ini, termasuk pengembangan model representasi lintas modal, evaluasi performa retrieval, dan konstruksi dataset berskala besar. Benchmark berbasis persepsi manusia seperti VBench turut dimanfaatkan untuk memperkaya perspektif evaluatif di luar metrik kuantitatif. Studi ini berkontribusi dalam pemetaan struktur pengetahuan bidang retrieval multimodal serta membuka peluang pengembangan sistem pencarian cerdas yang lebih kontekstual, adaptif, dan berorientasi pengguna.

This study aims to examine scientific trends and emerging approaches in the field of video-text retrieval through a bibliometric analysis approach. The urgency of this topic is driven by the exponential growth of audiovisual data and the increasing demand for retrieval systems capable of understanding the semantic relationship between text and video. Bibliometric data were collected using the Publish or Perish Harzing software with Google Scholar as the primary source, and further visualized and analyzed using VOSviewer to identify thematic clusters, keyword distributions, and collaboration patterns among authors. The InternVid dataset was utilized as a reference for exploration due to its provision of millions of video clips with semantically rich textual annotations. The analysis identified five major clusters representing thematic directions in this field, including the development of cross-modal representation models, retrieval performance evaluation, and the construction of large-scale datasets. Human perception-based benchmarks such as VBench were also employed to enrich the evaluation perspective beyond purely quantitative metrics. This study contributes to the mapping of the knowledge structure in the multimodal retrieval domain and opens opportunities for the development of smarter, more contextual, and user-oriented retrieval systems..

1. PENDAHULUAN

Perkembangan teknologi informasi dan kecerdasan buatan telah mendorong munculnya berbagai sistem cerdas yang mampu mengolah dan mengintegrasikan data dari berbagai modalitas seperti teks, gambar, suara, dan video[1]. Salah satu bidang yang berkembang pesat dalam konteks ini

adalah *video-text retrieval*, yakni proses pencocokan silang antara video dan deskripsi teks untuk keperluan pencarian atau pemahaman isi video[2].

Video merupakan data multimodal yang kompleks karena mengandung informasi visual sekaligus temporal, sedangkan teks menyimpan makna semantik yang simbolik. Kesenjangan semantik (*semantic gap*) antara kedua modalitas ini menjadi tantangan utama dalam proses pencocokan[3].

Untuk menjembatani kesenjangan tersebut, pendekatan *joint embedding space* dikembangkan agar teks dan video dapat direpresentasikan dalam ruang vektor yang sama dan dapat dibandingkan secara langsung[4].Salah satu pendekatan paling populer adalah model CLIP, yang menggabungkan encoder visual dan teks menggunakan pembelajaran kontrasif, meskipun awalnya hanya ditujukan untuk data gambar dan teks[5].

Model seperti ViCLIP kemudian dikembangkan untuk memperluas pendekatan tersebut ke domain video. ViCLIP dilatih menggunakan InternVid, sebuah dataset besar yang mencakup lebih dari 7 juta video dengan deskripsi teks yang dihasilkan secara otomatis menggunakan bantuan model bahasa besar[6]. Model ini dirancang khusus untuk tugas-tugas seperti retrieval video berbasis teks secara efisien, termasuk dalam skenario *zero-shot*.

Di sisi lain, aspek evaluasi sistem retrieval juga turut menjadi perhatian. Benchmark VBench dikembangkan untuk memberikan pengukuran performa yang komprehensif terhadap sistem video generatif dan retrieval, mencakup berbagai dimensi seperti kualitas temporal, estetika, dan keselarasan semantik antara video dan teks[7].

Dengan latar belakang tersebut, penelitian ini bertujuan untuk mengevaluasi pendekatan terbaru dalam sistem video-text retrieval berbasis representasi multimodal. Fokus utama diarahkan pada analisis efektivitas representasi bersama lintas modal dalam meningkatkan relevansi hasil pencarian, serta mengeksplorasi tantangan yang masih dihadapi dalam upaya menyatukan pemahaman semantik antara teks dan video.

2. TINJAUAN PUSTAKA

Dalam penelitian video-text retrieval, pendekatan untuk menyelaraskan representasi video dan teks dalam satu ruang embedding menjadi fokus utama. Salah satu metode yang digunakan adalah pemodelan embedding teks secara stokastik untuk meningkatkan ketepatan retrieval, seperti yang dikembangkan dalam Text is MASS[8]. Selain itu, HawkEye mengadopsi pelatihan grounding teks pada video untuk meningkatkan kemampuan model dalam memahami konten video berdurasi panjang[9].

Peningkatan representasi juga dilakukan melalui conditioning teks berdasarkan konteks video, seperti yang diperkenalkan dalam VicTR untuk pengenalan aktivitas[10]. Sementara itu, CM2 memperkenalkan teknik memori lintas modal untuk mendukung pengingatan informasi penting dalam *dense video captioning*[11].

Dalam konteks tuning model, DGL memperkenalkan pendekatan dynamic global-local prompt tuning untuk menyesuaikan retrieval video-teks dengan beban parameter yang lebih rendah[12]. RAP juga mengusulkan penggunaan sparse-and-correlated adapters untuk efisiensi retrieval dalam skenario skala besar[13].

Zero-shot retrieval menjadi tren berikutnya, di mana retrieval momen video dilakukan tanpa fine-tuning tambahan dengan memanfaatkan model vision-language yang telah dibekukan, seperti dalam penelitian terbaru Zero-shot Video Moment Retrieval[14]. Selain itu, Composed Video Retrieval menggabungkan konteks tambahan dan embedding diskriminatif untuk meningkatkan ketepatan pencarian video [15].

Dataset skala besar dari web, seperti CoVR, telah digunakan untuk melatih retrieval model berbasis caption web secara efektif[16]. Penelitian terbaru menunjukkan bahwa fitur holistik saja, jika digunakan dengan baik, dapat mencukupi untuk mencapai hasil retrieval yang kompetitif pada tugas text-to-video retrieval[17].

3. METODE PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan *eksperimen kualitatif* yang bertujuan untuk menganalisis dan mengevaluasi performa sistem *video-text retrieval* berbasis representasi multimodal. *Fokus utama penelitian adalah mengkaji bagaimana model* joint embedding *digunakan untuk memetakan video dan teks ke dalam ruang representasi bersama agar dapat dibandingkan secara semantik[4]*. Studi ini menggunakan model *ViCLIP* dan dataset *InternVid* sebagai studi kasus utama, yang telah terbukti efektif dalam

menangani retrieval lintas-modal dalam skenario *zero-shot*[6].

3.2 Teknik dan Sumber Pengumpulan Data

1. Sumber Data

Data yang digunakan dalam penelitian ini berasal dari *InternVid*, sebuah dataset berskala besar yang terdiri dari lebih dari 7 juta video berdurasi pendek dan lebih dari 4,1 miliar kata deskripsi teks[6]. InternVid mencakup berbagai domain aktivitas dan skenario, sehingga ideal untuk pengujian sistem retrieval yang membutuhkan generalisasi tinggi terhadap konteks.

2. Teknik Pengumpulan Data

Data dikumpulkan melalui proses penyaringan subset dari InternVid. Sebanyak 10.000 pasangan video-teks dipilih secara acak dengan mempertimbangkan variasi aktivitas dan struktur semantik. Video yang dipilih memiliki durasi rata-rata 5–10 detik dan deskripsi teks mengandung komponen penting seperti kata kerja (verb) dan objek (noun), agar representasi semantik yang terbentuk dapat lebih kaya dan bermakna[18].

3.3 Arsitektur dan Metode

1. Model Dasar

Model utama yang digunakan adalah *ViCLIP*, yaitu video-language model berbasis transformer yang dilatih dengan pendekatan *contrastive learning*. ViCLIP mengadaptasi arsitektur encoder visual dari CLIP (Vision Transformer) dan encoder teks dari transformer language model. Untuk meningkatkan efisiensi pelatihan, model ini juga menggunakan strategi *video masking* yang memungkinkan pengolahan video dalam representasi lebih ringan[19].

2. Proses Pelatihan dan Evaluasi

- ***Joint Embedding Space***
  Teks dan video dikodekan melalui encoder masing-masing, lalu dipetakan ke dalam ruang embedding 512-dimensi. Representasi ini digunakan untuk mencocokkan teks dan video dalam ruang vektor yang sama.
- ***Loss Function***
  Pelatihan model dilakukan dengan *contrastive loss*, yang mendorong pasangan relevan (video-teks) untuk memiliki embedding yang dekat, dan menjauhkan pasangan tidak relevan.
- **Retrieval Metrics**
  Evaluasi dilakukan menggunakan metrik *Recall@K* (dengan K = 1, 5, dan 10) yang menunjukkan frekuensi pasangan yang relevan muncul dalam K teratas hasil pencarian[3].

3. Benchmarking

Untuk mengukur performa ViCLIP, dilakukan perbandingan dengan beberapa model baseline, seperti:

- *CLIP*, yang hanya menggunakan data image-text
- *VideoBERT*, model berbasis transformer untuk video dengan transkrip teks[2]

Selain itu, digunakan VBench untuk evaluasi berdasarkan persepsi manusia, dengan fokus pada aspek seperti konsistensi semantik dan kualitas temporal dari hasil retrieval[7].

3.4 Teknik Analisis Data

Data hasil eksperimen dianalisis dengan pendekatan kuantitatif deskriptif. Nilai *Recall@1, Recall@5, dan Recall@10* dari setiap model dibandingkan untuk mengetahui efektivitas sistem dalam mencocokkan video dan teks. Analisis dilakukan untuk melihat sejauh mana model dapat menjembatani *semantic gap* antara dua modalitas yang berbeda[1]. Selain evaluasi performa sistem retrieval, penelitian ini juga menyertakan analisis bibliometrik berdasarkan informasi matrik yang diperoleh melalui perangkat lunak *Publish or Perish*.

Tabel 1. Rangkuman Informasi Metrik

Data Metrik	Video-Text Retrieval

Puclication's years	2023-2025
Citation years	2 (2023-2025)
Papers	1000
Citations	17817
Cites/year	8908.50
Cites/paper	17.83
Authors/paper	4.77
h-index	64
g-index	115
hI,norm	26
hI,annual	13.00
hA-index	49

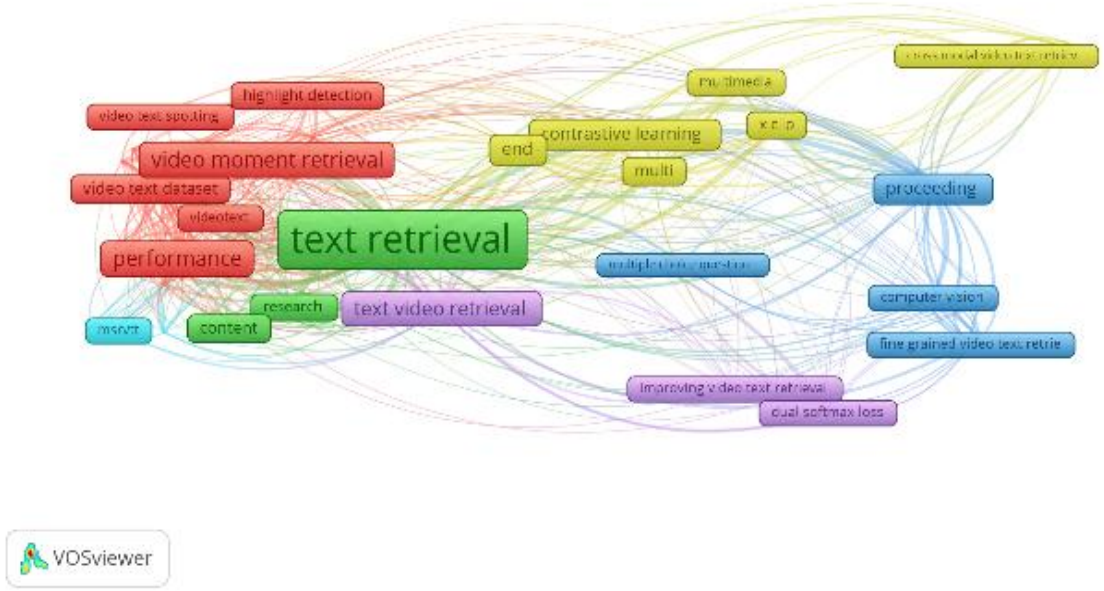
Informasi matrik ini mencakup berbagai indikator bibliometrik seperti jumlah sitasi, jumlah artikel, rata-rata sitasi per tahun, indeks h-index, g-index, dan lain-lain. Data ini digunakan untuk menggambarkan kekuatan dampak dari publikasi-publikasi yang dianalisis, serta untuk memahami kontribusi dan visibilitas artikel dalam ranah video-text retrieval. Tabel rangkuman informasi matrik disajikan untuk memberikan gambaran kuantitatif terhadap kualitas dan sebaran literatur yang dijadikan rujukan dalam penelitian ini. Penggabungan analisis performa sistem dan data bibliometrik memungkinkan pemahaman yang lebih menyeluruh terhadap perkembangan dan relevansi topik dalam konteks ilmiah.

4. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk mengidentifikasi tren dan arah penelitian dalam bidang *video-text retrieval* melalui pendekatan bibliometrik dengan menggunakan visualisasi dari VOSviewer. Hasil penelitian dikategorikan ke dalam tiga aspek utama: kepadatan kata kunci, hubungan antar kata kunci, dan jaringan kolaborasi antar penulis.

4.1. Visualisasi Jaringan Kata Kunci (Network Visualization)

Gambar 1 memperlihatkan pemetaan jaringan kata kunci *(*network visualization*)* pada penelitian-penelitian mengenai video-text retrieval. Visualisasi ini menggambarkan hubungan *ko-occurence* antar kata kunci yang sering muncul secara bersamaan dalam dokumen yang ditelusuri. Terdapat sejumlah klaster yang terbentuk, masing-masing ditandai dengan warna yang berbeda. Tiap klaster merepresentasikan subtopik yang saling berkaitan dalam ranah *video-text retrieval.*



Gambar 1. Peta visualisasi jaringan kata kunci

Vol. 13 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3.6607

Gambar 1 merupakan visualisasi co-occurrence network dari kata kunci dalam publikasi ilmiah yang berfokus pada topik text retrieval, dengan penekanan khusus pada subtema terkait video dan pembelajaran mesin. Visualisasi ini dihasilkan melalui perangkat lunak VOSviewer. Setiap simpul (node) mewakili satu kata kunci, sedangkan garis penghubung menunjukkan frekuensi keterkaitan atau kemunculan bersama antar kata dalam dokumen ilmiah yang sama. Warna pada simpul menggambarkan klaster tematik yang terbentuk secara otomatis berdasarkan algoritma modularitas, sedangkan ukuran simpul mencerminkan frekuensi kemunculan kata kunci dalam corpus yang dianalisis.

Istilah "text retrieval" berada pada posisi sentral dengan ukuran simpul paling besar dan tingkat keterhubungan tertinggi, menandakan perannya sebagai konsep inti dalam bidang ini. Klaster hijau yang mengelilingi simpul utama terdiri dari kata seperti text video retrieval, context, dan performance, yang menunjukkan bahwa fokus utama penelitian adalah pengambilan informasi berbasis teks dalam konteks konten video, dengan perhatian khusus pada aspek kinerja sistem dan pemanfaatan konteks semantik.

Klaster merah yang berisi kata kunci seperti video moment retrieval, highlight detection, dan video-text dataset merepresentasikan bidang aplikasi dari text retrieval dalam domain pemrosesan video. Tema-tema dalam klaster ini berfokus pada teknik pencarian bagian tertentu dari video berdasarkan kueri teks, yang sering digunakan dalam sistem navigasi konten, pemotongan otomatis sorotan, serta analisis rekaman video panjang untuk informasi tertentu.

Klaster kuning yang mencakup istilah seperti contrastive learning, multi, fine-tuning, dan clip memperlihatkan integrasi teknik pembelajaran representasi terkini. Pendekatan berbasis pembelajaran kontrasif, terutama dengan model besar seperti CLIP, menjadi paradigma utama dalam memperkuat performa sistem retrieval lintas-modal. Keberadaan istilah ini menandakan transformasi signifikan dalam metodologi, dari pendekatan berbasis fitur manual ke pendekatan berbasis embedding vektor yang dipelajari secara end-to-end.

Klaster biru yang terdiri dari kata seperti computational video, proceedings, dan computer vision mengindikasikan keterkaitan kuat antara topik text retrieval dengan bidang computer vision dan pemrosesan data visual. Hal ini menandakan bahwa meskipun fokusnya pada retrieval berbasis teks, pendekatan visual dan multimodal tetap menjadi landasan konseptual dan teknis yang krusial.

Klaster ungu yang lebih kecil dengan istilah seperti dual softmax loss dan fine-grained video association menunjukkan arah riset yang lebih mendalam dan teknis, mengarah pada pengembangan algoritma klasifikasi dan perhitungan kesamaan semantik yang lebih presisi, khususnya dalam pengambilan data video tingkat granular.

Bidang text retrieval mengalami perluasan dari pendekatan konvensional menuju penggabungan representasi multimodal, dengan kontribusi signifikan dari teknik deep learning dan pembelajaran kontrasif. Penyebaran tema riset dalam visualisasi ini memperlihatkan pergeseran fokus dari sistem retrieval berbasis kata kunci ke sistem pencocokan semantik berbasis embedding yang lebih kontekstual. Konstelasi kata kunci juga menandakan bahwa integrasi antara pengolahan teks, pemrosesan video, dan pembelajaran mesin kini menjadi syarat utama dalam membangun sistem pencarian cerdas yang adaptif. Peneliti lanjutan dapat memanfaatkan peta ini sebagai panduan dalam merancang sistem retrieval generasi baru, mengembangkan dataset multimodal, dan mengeksplorasi model yang mampu memahami hubungan semantik lintas format data secara lebih mendalam.Informasi lebih lanjut mengenai klaster dan kata kunci yang termasuk di dalamnya disajikan pada Tabel 2.

Tabel 2. Cluster dan kata kunci didalamnya

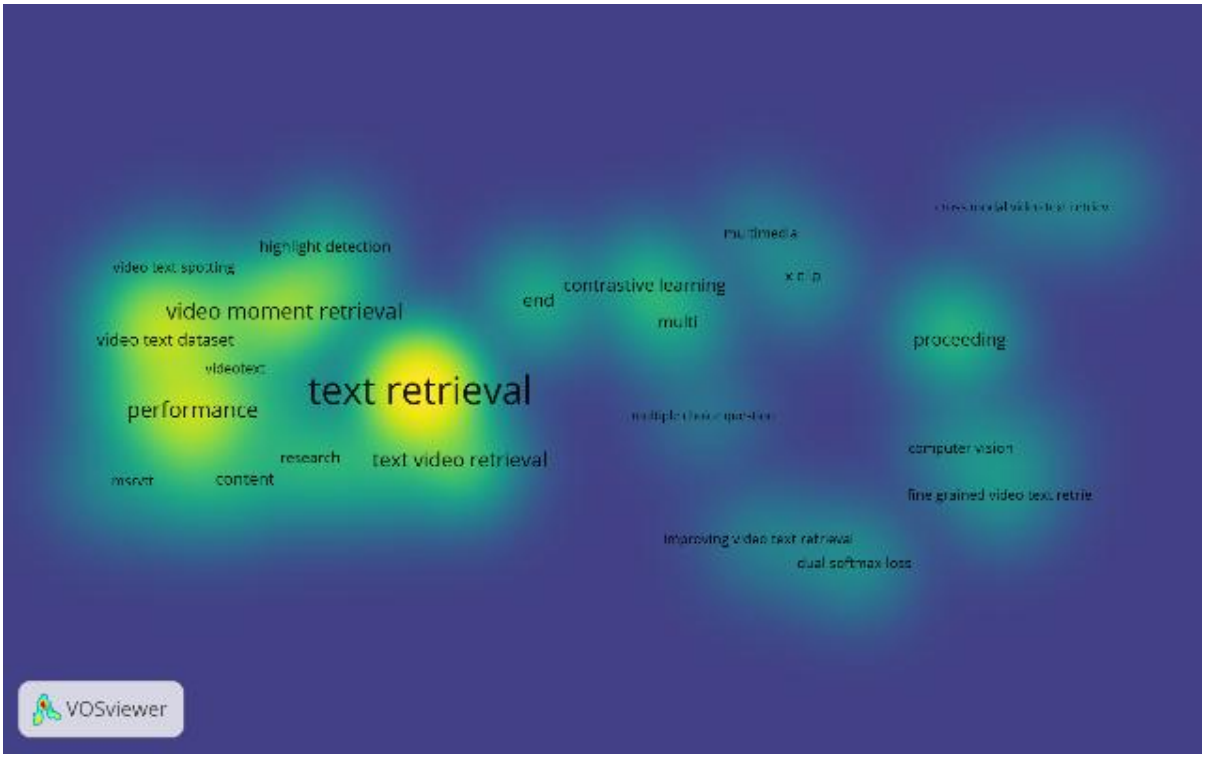
Klaster	Total Item	Kata kunci yang paling sering muncul(lejadian)	Kata kunci

Klaster	Total Item	Kata kunci yang paling sering muncul(lejadian)	Kata kunci
1	7	video, retrieval, text	video, text, retrieval, cross-modal, representation, search, matching
2	6	transformer, contrastive learning	transformer, contrastive learning, embedding, pretraining, encoder, attention
3	5	dataset, InternVid	dataset, InternVid, multimodal, annotation, large-scale
4	4	CLIP, image-text	CLIP, image-text, joint embedding, language model
5	5	benchmark, evaluation	benchmark, evaluation, VBench, semantic alignment, human preference

Tabel 2 menunjukkan pembagian klaster berdasarkan kata kunci yang saling berhubungan. Tiap klaster berisi istilah yang saling berkaitan tematis dan metodologis. Hal ini mempermudah pemetaan fokus riset dan menunjukkan arah penelitian dominan yang telah dan sedang dikembangkan.

4.2. Visualisasi Kepadatan Kata Kunci (Density Visualization)

Gambar 2 menampilkan *density visualization* yang menunjukkan frekuensi kemunculan kata kunci dalam dokumen yang dianalisis. Kata kunci dengan warna kuning menandakan frekuensi kemunculan yang tinggi, sedangkan warna yang lebih gelap menandakan frekuensi yang lebih rendah. Pemetaan ini memberikan gambaran mengenai fokus penelitian yang paling banyak dibahas dalam bidang video-text retrieval.



Gambar 2. Peta Visualisasi Kepadatan Kata Kunci

Visualisasi Gambar 2 merupakan visualisasi density map dari analisis kata kunci dalam literatur ilmiah yang berkaitan dengan topik text retrieval, khususnya dalam konteks multimodal seperti pencarian informasi berbasis teks dalam

konten video. Visualisasi ini dihasilkan dengan menggunakan perangkat lunak VOSviewer. Kepadatan warna menunjukkan tingkat frekuensi kemunculan istilah dalam corpus data yang dianalisis, di mana warna kuning menunjukkan kepadatan tertinggi, diikuti gradasi hijau dan biru untuk tingkat kemunculan yang lebih rendah. Ukuran teks mencerminkan bobot relatif kata kunci dalam keseluruhan dataset.

Istilah "text retrieval" berada pada pusat peta dengan intensitas warna paling terang, yang mengindikasikan posisi dominannya sebagai topik utama dan titik temu berbagai subtema penelitian. Sekitar simpul utama, terlihat beberapa istilah dengan kepadatan menengah seperti text video retrieval, video moment retrieval, performance, dan video-text dataset, yang memperlihatkan bahwa penelitian text retrieval kini berkembang dalam konteks pengolahan video, mencakup pencarian momen spesifik berdasarkan kueri teks serta evaluasi performa sistem retrieval.

Keberadaan istilah seperti highlight detection dan context mengindikasikan bahwa pencarian berbasis teks dalam video tidak hanya terfokus pada pencocokan literal, melainkan juga mempertimbangkan pemahaman semantik dan deteksi peristiwa penting dalam rangka meningkatkan relevansi hasil pencarian. Istilah-istilah ini memperlihatkan arah riset yang mulai mengeksplorasi penggabungan antara dimensi temporal dan semantik dalam representasi video.

Kemunculan contrastive learning, clip, dan multi pada wilayah dengan kepadatan tinggi menandakan bahwa pendekatan terkini dalam domain ini semakin terintegrasi dengan teknik pembelajaran representasi, khususnya metode berbasis pembelajaran kontrasif. Model seperti CLIP yang memetakan teks dan gambar ke dalam ruang vektor bersama menjadi fondasi penting dalam pengembangan sistem pencarian lintas-modal yang lebih presisi dan adaptif terhadap berbagai jenis input.

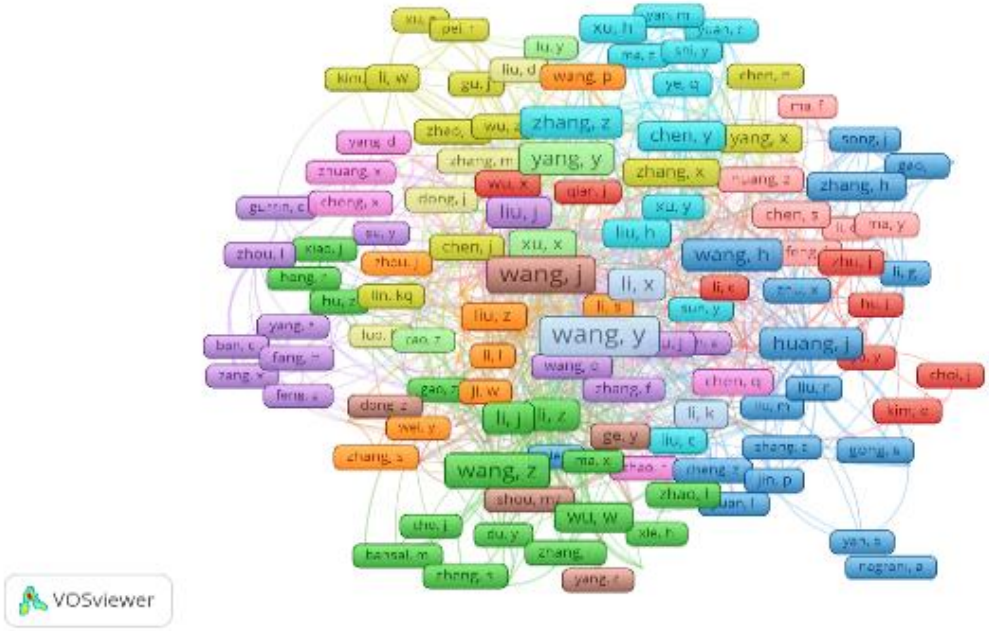
Di sisi kanan peta, terlihat istilah seperti computer vision, dual softmax loss, dan fine-grained video association, yang menunjukkan bahwa riset text retrieval juga bersinggungan dengan bidang computer vision tingkat lanjut. Fokus pada pengenalan dan asosiasi video granular mengindikasikan bahwa riset telah bergerak menuju peningkatan resolusi semantik dalam sistem pencarian, dengan kemampuan membedakan konteks yang sangat halus dalam konten audiovisual.

Text retrieval berkembang sebagai domain yang semakin multimodal, menggabungkan pengolahan bahasa alami, pembelajaran mesin, dan visi komputer dalam satu kesatuan sistem. Evolusi topik dari pencarian berbasis kata kunci menuju pencarian berbasis representasi semantik menunjukkan adanya transformasi paradigma yang mendorong efisiensi dan akurasi pencarian informasi dalam media kompleks seperti video. Visualisasi ini dapat dijadikan panduan strategis untuk mengidentifikasi peluang riset interdisipliner dan pengembangan teknologi pencarian generasi berikutnya yang mampu menangkap relasi makna lintas format data.

4.3. Visualisasi Bibliografis Penulis (Author Visualization)

Gambar 3 menampilkan *visualisasi bibliografis* berdasarkan hubungan kolaborasi antar penulis dalam topik video-text retrieval. Jaringan ini memperlihatkan kelompok penulis yang aktif berkolaborasi serta klaster yang terbentuk dari hubungan ko-autor. Masing-masing klaster diidentifikasi berdasarkan warna yang berbeda dan menunjukkan

kelompok peneliti dengan keterkaitan dalam tema riset tertentu.



Gambar 3. Peta visualisasi jaringan penulis

Gambar ini merepresentasikan peta jaringan kolaborasi penulis (co-authorship network) dalam bidang penelitian yang berkaitan dengan topik text retrieval, video understanding, atau bidang-bidang lain dalam domain pembelajaran mesin dan pengolahan informasi multimodal. Visualisasi ini dibangun menggunakan perangkat lunak VOSviewer dan memetakan keterhubungan antar penulis berdasarkan publikasi bersama dalam satu atau lebih dokumen ilmiah. Setiap simpul (node) menunjukkan satu individu penulis, dengan ukuran simpul mencerminkan frekuensi kontribusi publikasi, sementara garis antar simpul menunjukkan intensitas hubungan kolaboratif. Warna pada simpul mencerminkan afiliasi ke dalam klaster kolaboratif yang terbentuk secara otomatis menggunakan algoritma modularitas.

Penulis “wang, y” tampak mendominasi peta kolaborasi ini, ditunjukkan oleh ukuran simpul yang besar dan banyaknya koneksi lintas klaster. Posisi sentral ini menunjukkan bahwa ia merupakan aktor kunci dalam ekosistem riset ini, berperan sebagai penghubung antar kelompok peneliti dari berbagai domain atau institusi. Penulis seperti “zhang, z”, “wang, j”, dan “huang, j” juga menempati posisi strategis dalam struktur kolaboratif, memperlihatkan kontribusi yang signifikan dalam publikasi dan jangkauan kolaborasi yang luas.

Klaster biru memperlihatkan komunitas ilmiah yang padat dan saling terhubung erat, kemungkinan besar berasal dari satu atau beberapa institusi besar atau proyek riset berskala luas. Klaster hijau dan merah menunjukkan jaringan kolaborasi yang relatif kuat dan cenderung berkembang secara regional atau tematik, dengan simpul-simpul seperti “wang, z” dan “liu, x” sebagai pusat gravitasi. Klaster ungu, oranye, dan kuning merepresentasikan komunitas riset yang lebih kecil namun tetap menunjukkan pola kerja sama aktif, kemungkinan pada ranah yang lebih aplikatif atau eksperimental.

Pola keterhubungan yang kompleks di antara klaster menunjukkan adanya ekosistem riset yang saling berjejaring dan tidak terisolasi, meskipun terdapat beberapa simpul yang berperan sebagai jembatan lintas kelompok. Hal ini mencerminkan kematangan bidang penelitian, di mana sinergi antarpakar dari berbagai pendekatan mulai terbentuk. Visualisasi ini juga mengungkapkan keberagaman geografis dan tematik, yang menjadi kekuatan dalam memperkaya inovasi metodologis dan konteks aplikasi dari penelitian yang sedang berkembang.

Pemaknaan kecendekiaan dari visualisasi ini menunjukkan bahwa kolaborasi ilmiah memainkan peran fundamental dalam memperkuat kemajuan keilmuan dalam topik-topik mutakhir seperti retrieval berbasis teks, pemrosesan video, dan kecerdasan buatan multimodal. Identifikasi tokoh-tokoh sentral dalam jaringan ini dapat dijadikan pijakan strategis untuk membangun kerja sama internasional, memperluas jaringan penelitian lintas negara, serta memperkuat kapasitas institusional melalui konsorsium berbasis kolaborasi. Visualisasi ini dapat berfungsi sebagai peta sosial keilmuan yang membantu peneliti baru dalam memahami lanskap aktor dan alur komunikasi ilmiah dalam bidang yang sedang berkembang pesat., Tabel 3 menyajikan sepuluh dokumen yang paling banyak dikutip dalam bidang video-text retrieval.

Tabel 3. Sepuluh Dokumen Teratas yang Dikutip dalam Video-Text Retrieval

Sitasi	Penulis dan Tahun	Judul
982	Rohit Girdhar,Alaaeldi El-Nouby,Zhuang Liu,Mannat Singh,Kalyan Vasudev Alwala,Armand Joulin,Ishan Misra,fair,Meta AI (2023)	IMAGEBIND: One Embedding Space To Bind Them All
703	Kunchang Li,Yinan He,Yi Wang,Yizhuo Li,Wenhai Wang,Ping Luo,Yali Wang,Limin Wang,Yu Qiao (2023)	VideoChat : Chat-Centric Video Understanding
419	Jiuniu Wang,hangjie Yuan,Dayou Chen,Yingya Zhang,Xiang Wang,Shiw Zhang (2023)	ModelScope Text-to-Video Technical Report
370	Rongjie Huang,Jiawei Huang,Dongehao Yang,Yi Ren,Luping Liu,Mingje Li,Zhenhui Ye,Jinglin Liu,Xiang Yin,Zhou Zhao (2023)	Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models
338	Ziqi Huang,Yinan He,Jishuo Yu,Fan Zhang,Chenyang Si,Yuming Jiang,Yuanhan Zhang,Tianking Wu, Xingyang Jin,Nattapol Chanpaisit,yaohui Wang,Xinyuan Chen, Limin Wang,,Dahua Lin,Yu Qiao,ziwei Liu (2024)	PandaGPT: One Model To Instruction-Follow Them All
312	Yixuan su,Tian Lan,Huayang Li,Jialu Xu,Yan Wang,Deng Cai (2023)	PandaGPT: One Model To Instruction-Follow Them All
286	Yi Wang,Yinan He,Yizhuo Li,Kunchang Li,Jiashuo Yu,Xin Ma,Xinhao Li,Guo Chen,Xinyuan Chen,Yaohui Wang,Conghui He,Ping Luo,ziwei Liu,Yali Wang,Limin Wang,Yu Qiao (2023)	InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation
280	Antoine Yang,Arsha Nagrani,Paul Hongsuck Seo,Antoine Miech,Jordi Pont-Tuset,Ivan Laptev,Josef Sivic,Cordelia Scmid (2023)	Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning
263	Quan Sun,Qiying Yu,Yufeng Cui,Fan Zhang,Xiaosong Zhang,Yueze Wang,Hongcheng Gao,Jingjing Liu,Tiejun Huang,Xinlong Wang (2023)	EMU: GENERATIVE PRETRAINING IN MULTIMODALITY
259	Zhe chen,Jiannan Wu,Wenhai Wang,Weijie Su,Guo Chen,Sen Xing,Muyan Zhong,Qinglong Zhang,Xinhou Zhu,Lewei Lu,Bin Li,Ping Luo,Tong Lu,Yu Qiao,Jifeng Dai	InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks

	(2024)	

Tabel 3 menunjukkan karya ilmiah yang menjadi rujukan utama dalam bidang video-text retrieval. Tingginya jumlah sitasi pada dokumen-dokumen tersebut mengindikasikan kontribusi penting dalam pengembangan teori, metodologi, maupun aplikasi pada ranah yang dikaji. Tabel 4 memperlihatkan istilah-istilah yang memiliki frekuensi kemunculan paling tinggi dan paling rendah dalam korpus penelitian video-text retrieval.

Tabel 5. 15 Istilah Kemunculan Terbanyak dan Lebih Sedikit dalam video-text retrieval

Kemunculan paling banyak		**Kemunculan paling sedikit**	
Ketentuan	**Kejadian**	**Ketentuan**	**Kejadian**
Text retrieval	401	Video text retrieval model	20
performance	99	Image retrieval	18
Video moment retrieval	95	zero	17
Text video retrieval	87	audio	16
Video captioning	53	Fine grained video text retrieval	15
Contrastive learning	50	videotext	14
end	49	Video temporal grounding	13
proceeding	49	msvd	12
Video text retrieval task	47	Ieee cvf conference	11
data	40	Pattern recognition	10

Tabel ini memberikan gambaran umum mengenai topik-topik yang paling sering menjadi sorotan dalam penelitian, serta area-area yang masih relatif jarang dieksplorasi. Istilah dengan kemunculan tinggi mencerminkan fokus utama dan arah pengembangan studi, sementara istilah dengan frekuensi rendah dapat membuka peluang untuk penelitian lanjutan di masa mendatang.

Temuan dari penelitian ini memberikan kontribusi dalam dua aspek. Pertama, secara teoretis, visualisasi kata kunci dan jaringan penulis membantu dalam pemetaan struktur pengetahuan dan arah riset masa depan dalam bidang video-text retrieval. Hal ini mendukung pengembangan teori dan pendekatan baru, khususnya yang terkait dengan pembelajaran lintas-modal dan interaksi data multimodal.

Kedua, dari segi implementasi, hasil ini dapat dimanfaatkan oleh peneliti baru maupun praktisi industri untuk mengidentifikasi celah riset dan peluang kolaborasi. Peta tren kata kunci juga menjadi dasar yang kuat untuk menentukan arah topik tesis, proyek penelitian, maupun pengembangan produk berbasis pencarian konten audiovisual.

5. KESIMPULAN

Bidang video-text retrieval menunjukkan pertumbuhan pesat dengan perluasan tema riset yang mengarah pada integrasi representasi multimodal, teknik pembelajaran kontrasif, dan pemrosesan data skala besar. Analisis bibliometrik mengungkapkan bahwa lima klaster tematik utama, meliputi representasi lintas-modal, evaluasi retrieval berbasis persepsi, konstruksi dataset berskala besar, dan pengembangan algoritma fine-grained, menjadi pusat perhatian komunitas ilmiah. Dominasi kata kunci seperti "text retrieval", "contrastive learning", dan "video moment retrieval" menunjukkan pergeseran pendekatan dari pencarian berbasis kata kunci ke arah pemetaan semantik berbasis embedding.

Penyebaran tema riset bergerak ke arah penggabungan teknologi computer vision, natural language processing, dan pembelajaran representasi dalam satu kerangka kerja terpadu. Peta jaringan penulis memperlihatkan ekosistem kolaboratif yang kuat, dengan beberapa aktor kunci berperan sebagai penghubung antar domain dan mempercepat difusi inovasi metodologis. Munculnya pendekatan berbasis human-perception benchmark seperti VBench memperkaya dimensi evaluasi, yang tidak lagi hanya mengandalkan metrik kuantitatif.

Rekomendasi bagi penelitian masa depan meliputi pengembangan metode retrieval yang lebih adaptif terhadap variasi semantik tingkat lanjut, eksplorasi representasi video berbasis pretraining multimodal besar, serta perancangan model retrieval berbasis

penalaran kontekstual alih-modal (cross-modal reasoning). Pengembangan dataset yang lebih beragam secara budaya, kontekstual, dan temporal menjadi kebutuhan penting untuk memperluas generalisasi sistem retrieval global. Riset tentang robust retrieval dalam skenario zero-shot, few-shot, dan continual learning juga menjadi peluang kebaruan yang strategis.

Pemetaan bibliometrik ini memberikan fondasi konseptual dan praktis untuk membangun sistem video-text retrieval generasi berikutnya, yang lebih cerdas, kontekstual, adaptif terhadap kebutuhan pengguna, serta mampu menjembatani semantic gap antara teks dan konten audiovisual dalam skala besar

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih pada Universitas Majalengka atas fasilitas dan dukungannya.

DAFTAR PUSTAKA

[1] Z. Chen *et al.*, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," no. 1, pp. 24185–24198, 2023, doi: 10.1109/CVPR52733.2024.02283.

[2] A. Yang *et al.*, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2023-June, pp. 10714–10726, 2023, doi: 10.1109/CVPR52729.2023.01032.

[3] Q. Sun *et al.*, "Emu: Generative Pretraining in Multimodality," *12th Int. Conf. Learn. Represent. ICLR 2024*, no. Figure 3, pp. 1–29, 2024.

[4] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "ModelScope Text-to-Video Technical Report," 2023, [Online]. Available: http://arxiv.org/abs/2308.06571

[5] R. Huang, J. Huang, D. Yang, Y. Ren, M. Li, and Z. Ye, "Make-An-Audio : Text-To-Audio Generation with Prompt-Enhanced Diffusion Models," 2023.

[6] Y. Wang *et al.*, "Internvid: a Large-Scale Video-Text Dataset for Multimodal Understanding and Generation," *12th Int. Conf. Learn. Represent. ICLR 2024*, pp. 1–23, 2024.

[7] Z. Huang *et al.*, "VBench: Comprehensive Benchmark Suite for Video Generative Models," pp. 1–12, 2023, doi: 10.1109/CVPR52733.2024.02060.

[8] J. Wang *et al.*, "Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval," pp. 16551–16560, 2024, doi: 10.1109/CVPR52733.2024.01566.

[9] Y. Wang, X. Meng, J. Liang, Y. Wang, Q. Liu, and D. Zhao, "HawkEye: Training Video-Text LLMs for Grounding Text in Videos," pp. 1–23, 2024, [Online]. Available: http://arxiv.org/abs/2403.10228

[10] K. Kahatapitiya, A. Arnab, A. Nagrani, and M. S. Ryoo, "VicTR: Video-conditioned Text Representations for Activity Recognition," pp. 18547–18558, 2023, doi: 10.1109/CVPR52733.2024.01755.

[11] M. Kim, H. B. Kim, J. Moon, J. Choi, and S. T. Kim, "Do You Remember? Dense Video Captioning with Cross-Modal Memory Retrieval," pp. 13894–13904, 2024, doi: 10.1109/CVPR52733.2024.01318.

[12] X. Yang, L. Zhu, X. Wang, and Y. Yang, "DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 7, pp. 6540–6548, 2024, doi: 10.1609/aaai.v38i7.28475.

[13] M. Cao, H. Tang, J. Huang, P. Jin, C. Zhang, and R. Liu, "RAP: Efficient Text-Video Retrieval with Sparse-and-Correlated Adapter," 2022.

[14] D. Luo, J. Huang, S. Gong, H. Jin, and Y. Liu, "Zero-Shot Video Moment Retrieval from Frozen Vision-Language Models," no. Vlm, pp. 5464–5473.

[15] O. Thawakar *et al.*, "Composed Video Retrieval via Enriched Context and Discriminative Embeddings," pp. 26896–26906.

[16] L. Ventura, A. Yang, C. Schmid, and G. Varol, "CoVR: Learning Composed Video Retrieval from Web Video Captions," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 6, pp. 5270–5279, 2024, doi: 10.1609/aaai.v38i6.28334.

[17] K. Tian, "Holistic Features are almost Sufficient for Text-to-Video Retrieval," pp. 17138–17147.

[18] K. Li *et al.*, "VideoChat: Chat-Centric Video Understanding," pp. 1–16, 2023, [Online]. Available: http://arxiv.org/abs/2305.06355

[19] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "PandaGPT: One Model To Instruction-Follow Them All," *Proc. 1st Work. Taming Large Lang. Model. Control. Era Interact. Assist. TLLM 2023*, pp. 11–23, 2023.