INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 11 No. 1 Juni 2025, pp. 95-107

PERSPEKTIF GLOBAL TREN DAN PERKEMBANGAN INOVASI PENELITIAN *VIDEO TO MUSIC GENERATION*

Ade Bastian[1], Ardi Mardiana[2], Muhammad Fahmi Ajiz[3] , Satria Winata[4]
[1,2,3,4]Program Studi Informatika, Fakultas Teknik, Universitas Majalengka
Email: fahmiajiz@unma.ac.id

ABSTRACT
This research aims to map the evolution of AI-based music generation, specifically focusing on music generation from video. Through bibliometric analysis of 999 scientific publications (1997-2025), we analyzed trends and conceptual structures using VOSviewer. Methods included metadata extraction, co-authorship network construction, and dominant cluster identification. Results revealed five major thematic clusters: text-based generative models, symbolic music generation, video game music, multimedia integration, and automatic composition. Recent studies show a shift toward multimodal generative architectures, integrating transformers and diffusion models to address semantic-temporal alignment challenges between video and music. The research identifies key gaps: scarcity of large-scale paired datasets, lack of standard evaluation metrics, and limited real-time generation systems. The novelty of this research lies in providing the first bibliometric mapping exclusively focused on music generation from video, offering a foundation for academic and industry communities to understand the trajectory and future directions of this rapidly evolving field.

Keywords: Artificial Intelligence, AI-Driven Music Generation, Bibliometric Analysis, Generative Models, Video to Music Generation

ABSTRAK
Penelitian ini bertujuan memetakan evolusi generasi musik berbasis AI, khususnya generasi musik dari video. Melalui analisis bibliometrik terhadap 999 publikasi ilmiah (1997-2025), kami menganalisis tren dan struktur konseptual menggunakan VOSviewer. Metode meliputi ekstraksi metadata, konstruksi jaringan ko-kepengarangan, dan identifikasi kluster dominan. Hasil mengungkapkan lima kluster tematik utama: model generatif berbasis teks, generasi musik simbolik, musik video game, integrasi multimedia, dan komposisi otomatis. Studi terbaru menunjukkan pergeseran ke arsitektur generatif multimodal, mengintegrasikan transformer dan model difusi untuk mengatasi tantangan penyelarasan semantik-temporal antara video dan musik. Penelitian mengidentifikasi kesenjangan utama: kelangkaan dataset berpasangan skala besar, kurangnya metrik evaluasi standar, dan terbatasnya sistem generasi real-time. Kebaruan penelitian ini adalah pemetaan bibliometrik pertama yang fokus eksklusif pada generasi musik dari video, memberikan fondasi bagi komunitas akademik dan industri untuk memahami lintasan dan arah masa depan bidang ini.

Kata Kunci: Artificial Intelligence, AI-Driven Music Generation, Bibliometric Analysis, Generative Models, Video to Music Generation

Riwayat Artikel :

Tanggal diterima : 14-05-2025

Tanggal revisi : 05-06-2025

Tanggal terbit : 20-06-2025

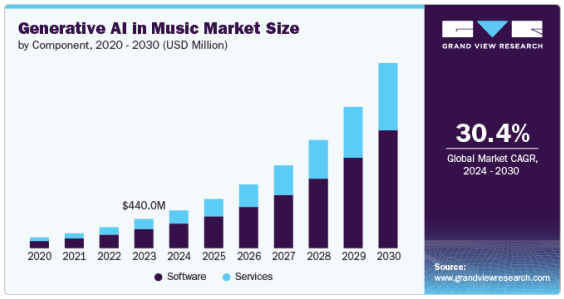
DOI :

https://doi.org/10.31949/infotech.v11i1.13830

1. PENDAHULUAN

1.1. Latar Belakang

Bidang video-to-music generation yaitu pembuatan musik secara otomatis yang selaras dengan konten video berkembang pesat seiring meningkatnya kebutuhan konten multimedia. Musik latar terbukti memainkan peran krusial dalam meningkatkan kualitas pengalaman menonton video, menarik perhatian audiens, dan memperkuat emosi atau narasi yang disampaikan(Ji et al., 2025). Platform berbagi video modern seperti YouTube dan TikTok, telah mendorong lonjakan konten video amatir, sehingga semakin banyak kreator yang membutuhkan musik tepat guna untuk melatari video mereka(Zhang & Fuentes, 2024). Proses konvensional mengandalkan kurasi manual oleh editor profesional atau pencarian musik bebas royalti, yang tidak hanya memakan waktu tetapi sering kali gagal menemukan musik yang benar-benar sesuai mood dan ritme video(Zhang & Fuentes, 2024). Signifikansi ekonomi dari teknologi ini tercermin dalam proyeksi pertumbuhan pasar Generative AI dalam industri musik yang diperkirakan mencapai CAGR 30,4% pada periode 2024-2030 (Gambar 1).



Gambar 1. Proyeksi pertumbuhan pasar Generative AI dalam industri musik 2020-2030. Sumber: Grand View Research

Jiang et al. (2017) menekankan bahwa video secara inheren bersifat multimodal, sehingga pemahaman menyeluruh tentang konten video, termasuk generasi musik, memerlukan pendekatan yang mengintegrasikan berbagai petunjuk modalitas seperti informasi spasial statis, pola gerakan, audio, serta dinamika temporal jangka panjang, yang justru menjadi fokus utama dalam pemetaan evolusi bibliometrik penelitian video-to-music generation(Jiang et al., 2017). Kesulitan ini kian diperparah oleh isu hak cipta dan keterbatasan stok musik, menjadikan otomatisasi video soundtrack sebagai kebutuhan mendesak(Zhang & Fuentes, 2024). Dengan kata lain, video-to-music generation muncul sebagai solusi menjanjikan untuk menghadirkan musik latar yang tailor-made secara cepat dan fleksibel bagi beragam video(Kang et al., 2024).

Eksplorasi awal mengenai hubungan antara konten visual dan audio telah dimulai sejak awal 2000-an. Pionir dalam bidang ini, Dannenberg dan Neuendorffer(Dannenberg & Neuendorffer, 2003) mengembangkan sistem yang memanfaatkan intensitas cahaya dalam video sebagai pengendali amplitudo harmonik, memungkinkan sintesis suara secara waktu nyata yang berubah sesuai gerakan visual dalam citra video. Pendekatan awal ini, meskipun masih sangat sederhana, telah membuka jalan untuk penelitian selanjutnya tentang korelasi audiovisual. Beberapa tahun kemudian, Zhou dkk.(Zhou et al., 2018) membawa konsep ini lebih jauh melalui pendekatan Visual to Sound yang berfokus pada sintesis suara alami dari video dunia nyata. Dengan memanfaatkan arsitektur encoder-decoder berbasis SampleRNN, model ini berhasil menghasilkan waveform mentah secara langsung dari input visual, menunjukkan bahwa korelasi antara gerakan visual dan suara dapat dipelajari bahkan dalam lingkungan video yang kompleks dan tidak terkontrol.

1.2. Tinjuan Pustaka

Mehri et al. (2017) mengembangkan SampleRNN, sebuah model generatif audio berbasis jaringan saraf yang menggunakan arsitektur hierarkis dengan modul yang beroperasi pada skala waktu berbeda, yang sangat relevan dalam konteks sintesis musik dari video dimana model perlu menangkap struktur temporal panjang sekaligus detail halus pada resolusi sampel tinggi, suatu pendekatan yang menjadi tren penting dalam evolusi generasi musik berbasis AI seperti ditunjukkan dalam analisis bibliometrik(Mehri et al., 2017).

Berbagai pendekatan telah dikembangkan dalam upaya mengatasi tantangan video-to-music generation, dengan keunggulan dan keterbatasan yang berbeda-beda. Tabel 1 menyajikan komparasi pendekatan utama yang telah dikembangkan dalam dua dekade terakhir, mengidentifikasi kesenjangan yang masih perlu diatasi dalam penelitian ke depan.

Tabel 1. Komparasi Pendekatan Utama dalam Video-to-Music Generation

Peneliti dan Tahun	Pendekatan/Representasi	Kekuatan & Keterbatasan Utama
Dannenberg & Neuendorffer (2003)	Reaktivitas visual → Sintesis parametrik	(+) Pionir korelasi visual-audio (−) Tanpa pemahaman semantik, fitur tingkat rendah
Zhou et al. (2018)	Visual to Sound (SampleRNN) → Waveform	(+) Sintesis langsung dari video dunia nyata

Peneliti dan Tahun	Pendekatan/Representasi	Kekuatan & Keterbatasan Utama
		(−) Terbatas pada suara natural, bukan musik terstruktur
Di et al. (2021)	Controllable Music Transformer → MIDI simbolik	(+) Kendali eksplisit fitur video (tempo, gerakan) (−) Rule-based, tidak belajar dari data tersinkron
Zhuo et al. (2023)	Video Background Music → Piano-roll + diffusion	(+) Dataset SymMV, metrik VMCP, multimodal (−) Struktur musik jangka panjang terbatas
Su et al. (2024)	V2Meow (U-Net diffusion) → Waveform	(+) Kualitas audio tinggi, prompt teks untuk control (−) Koherensi struktural musik lemah
Li et al. (2024)	Diff-BGM → Piano-roll → audio	(+) Segment-aware diffusion, polyphonic music (−) Ketergantungan pada segmentasi video
Lin et al. (2024)	VMAS → Simbolik + audio	(+) Video-beat alignment, data web musik besar (−) Bias terhadap gaya musik populer
Rai & Sridhar (2024)	EgoSonics → Latent diffusion	(+) Sinkronisasi temporal presisi dengan visual (−) Domain terbatas (video egosentris)

Meskipun perkembangan awal menjanjikan, tugas video-to-music generation menghadirkan tantangan multidimensi yang signifikan. Tantangan utama terletak pada kesenjangan modalitas antara domain visual dan musikal(Gu et al., 2023). Konten visual membawa informasi semantik (objek, adegan) dan dinamika temporal (pergerakan, transisi adegan) yang harus diterjemahkan menjadi elemen musikal seperti melodi, harmoni, dan ritme(S. Li, Qin, et al., 2024). Hubungan antara video dan musik tidak bersifat deterministik satu-satu, melainkan dipengaruhi gaya estetika dan emosi(Zhuo et al., 2023). Video dengan perubahan adegan cepat, misalnya, memerlukan musik berirama cepat dan enerjik, sedangkan video bernuansa tenang cocok dipadukan dengan musik lambat nan menenangkan(S. Li, Qin, et al., 2024). Demikian pula, klip gembira seharusnya diiringi melodi ceria, sementara adegan sedih lebih selaras dengan musik bernuansa minor yang mendalam(S. Li, Qin, et al., 2024). Mengintegrasikan korespondensi semantik-emosional semacam ini ke dalam model generatif merupakan tantangan tersendiri(Zhuo et al., 2023).

Berbeda dengan penelitian sebelumnya, Wang et al. (2025) menegaskan bahwa meskipun video-to-music generation memiliki potensi aplikasi yang luas dalam scoring film, platform video pendek, dan sintesis musik tarian, penelitian di bidang ini masih berada dalam tahap awal perkembangannya karena kompleksitas struktur internal musik dan tantangan dalam pemodelan hubungan dinamis dengan video(Wang et al., 2025). Penelitian terdahulu bahkan menyebut adanya asimetri informasi: konten visual sangat beragam sementara informasi emosional musik sangat kaya, sehingga menyulitkan pemetaan keduanya secara langsung(Gu et al., 2023).

Akibatnya, banyak pendekatan awal hanya memanfaatkan korelasi fitur tingkat rendah (mis. kesamaan pola warna atau gerak) dan mengabaikan konteks semantik yang lebih tinggi(J.-C. Lin et al., 2016). Lin dkk. (2016) menyoroti bahwa tanpa mempertimbangkan semantik (seperti emosi "sedih" vs "gembira"), sulit menjembatani modalitas musik dan video hanya dari ciri low-level(J.-C. Lin et al., 2016). Tantangan berikutnya adalah sinkronisasi waktu: musik harus selaras secara temporal dengan video, misalnya perubahan dramatis pada video sebaiknya diiringi aksen atau perubahan dinamika musik(Prétet, 2022). Penyelarasan ritmik ini sangat sulit dicapai melalui metode generatif naïf, sehingga menjadi hambatan perkembangan model-model awal(Kang et al., 2024). Selain itu, musik memiliki struktur internal kompleks (frasa, progresi akor, dll.) yang harus dijaga koherensinya; menciptakan musik yang struktural sekaligus mengikuti ritme video merupakan masalah terbuka dalam generasi musik AI(Briot et al., 2019).

Disamping tantangan teknis, keterbatasan data dan evaluasi turut menghambat kemajuan video-to-music generation. Berbeda dari text-to-music atau speech-to-music yang belakangan ini didukung kumpulan data besar, hingga beberapa tahun lalu hampir tidak ada dataset video-musik berpasangan yang memadai(Zhang & Fuentes, 2024). Data video dengan musik latar sering kali sulit diperoleh karena

isu lisensi/copyright dan keharusan sinkronisasi yang presisi(Zhang & Fuentes, 2024). Akibatnya, beberapa studi awal terpaksa menggunakan data terbatas seperti video musik resmi (yang cenderung tersinkron secara artistik namun domainnya sempit) atau bahkan data tak berpasangan(Su et al., 2024). Zhuo dkk. (2023) menggarisbawahi ketiadaan dataset skala besar yang memuat video berkualitas dan musik teranotasi secara selaras sebagai kendala utama pertama di bidang ini(Zhuo et al., 2023). Tanpa data memadai, model sulit belajar hubungan lintas-modal secara general. Untuk mengatasi ini, penelitian terbaru mulai menyusun dataset khusus: misalnya SymMV yang memuat 1.140 pasangan video dan musik simbolik (MIDI) berbagai genre beserta anotasi kord, melodi, dan iringan(Zhuo et al., 2023). Langkah ini penting karena representasi musik simbolik mempermudah model menangkap struktur dan features semantik musikal(Gan et al., 2020).

Di lain pihak, keterbatasan metode evaluasi juga menjadi masalah krusial. Penilaian kecocokan musik-video kerap bersifat subjektif, mengandalkan uji pendengar manusia(Zuo et al., 2025). Survei menunjukkan banyak karya terdahulu mengevaluasi hasil dengan user study atau metrik musik generik yang tidak sensitif terhadap keselarasan video(Zuo et al., 2025). Hal ini tidak hanya mahal dan bias, namun juga tidak memberikan umpan balik yang terstruktur bagi optimisasi model(Zuo et al., 2025). Para peneliti telah mengakui kurangnya metrik objektif untuk menilai korespondensi video-musik sebagai tantangan utama ketiga bidang ini(Zhuo et al., 2023). Upaya menuju solusi mencakup rancangan metrik berbasis retrieval seperti VMCP (Video-Music CLIP Precision) yang menggunakan model joint embedding video-musik untuk mengukur seberapa tepat musik mengiringi video(Zhuo et al., 2023). Pendekatan metrik baru ini diharapkan mampu mengevaluasi alignment tempo, mood, dan dinamika secara lebih kuantitatif dibanding sekadar penilaian subjektif.

Dalam konteks retrieval video-to-music, Stewart et al. (2024) mengembangkan pendekatan inovatif berbasis pembelajaran kontrastif semi-supervised yang tidak hanya mampu menemukan musik yang sesuai untuk video tertentu, tetapi juga menawarkan kontrol eksplisit terhadap seberapa besar sistem fokus pada informasi yang dipelajari secara mandiri (self-supervised) versus informasi berlabel, sehingga memungkinkan penyesuaian dinamis antara kesesuaian audio-visual alami dan pengetahuan domain spesifik(Stewart et al., 2024).

Sejalan dengan tantangan di atas, komunitas riset telah mengusulkan berbagai pendekatan teknologi yang inovatif. Perkembangan deep learning sangat berperan dalam mendorong batas generative AI lintas modal(Božić & Horvat, 2024). Secara garis besar, pendekatan video-to-music generation dapat dibedakan berdasarkan representasi musik yang digunakan (simbolik vs audio mentah) serta jenis data latih (berpasangan vs tak berpasangan), di samping arsitektur model generatif yang dipakai (mis. transformer, GAN, diffusion).

Pendekatan berbasis representasi simbolik (seperti MIDI) banyak diadopsi pada penelitian awal(Gan et al., 2020). Representasi ini menyimpan notasi musik (nada, durasi, dinamika) secara terstruktur, sehingga memudahkan kontrol dan interpretabilitas oleh model(Gan et al., 2020). Chen dkk. (2020) misalnya menggunakan MIDI sebagai target output agar model belajar hubungan gerakan tubuh pemain instrumen dengan nada yang dihasilkan(Gan et al., 2020). Dengan kerangka Graph-Transformer, model tersebut berhasil memetakan gerak tangan pada video menjadi rangkaian nada MIDI, kemudian disintesis menjadi audio(Gan et al., 2020). Keunggulan MIDI adalah menjaga kejelasan struktur musik (progresi akor, pola ritme) sehingga musik hasil generatif lebih koheren dan dapat diedit lanjut oleh komposer di Digital Audio Workstation(Suríš et al., 2022). Namun, Lin dkk.(Y.-B. Lin et al., 2024) menunjukkan bahwa pendekatan konvensional berbasis MIDI memiliki keterbatasan ekspresivitas. Untuk mengatasi hal ini, mereka memperkenalkan VMAS (Video-Music Alignment Scheme) yang memanfaatkan data besar dari video musik web untuk melatih model generatif Transformer. Model ini menyelaraskan musik secara semantik dan ritmis dengan video input melalui kombinasi pelatihan autoregressive dan contrastive learning, serta skema video-beat alignment untuk memastikan sinkronisasi yang presisi dengan dinamika visual.

Liu et al. (2025) mengusulkan DyViM, sebuah kerangka kerja inovatif untuk generasi musik dari video yang menitikberatkan pada pemodelan dinamika visual melalui ekstraksi fitur dinamika frame-wise dari metode berbasis optical flow, serta penyelarasan temporal tingkat token untuk mengatasi ketidaksesuaian representasi antara video dan musik, suatu pendekatan yang mencerminkan perkembangan terkini dalam evolusi algoritma generasi musik berbasis AI yang terpetakan dalam analisis bibliometrik(Liu et al., 2025).

Sebaliknya, pendekatan yang menghasilkan audio mentah (waveform) langsung berusaha menciptakan musik dengan warna bunyi lengkap. Model diffusion dan generative transformer mutakhir memungkinkan generasi audio resolusi tinggi dari kondisi visual(Su et al., 2024). Contohnya, model V2Meow (AAAI 2024) melatih U-Net diffusion untuk menyintesis waveform musik dari video input(Su et al., 2024). Model ini bahkan mendukung prompt teks untuk mengendalikan gaya musik (seperti memilih instrumen atau genre)(Su et al., 2024). Hasil V2Meow menunjukkan fidelitas audio yang baik, namun pendekatan ini menghadapi tantangan dalam menjaga koherensi struktur musikal jangka panjang(Su et al., 2024). Para peneliti mencatat bahwa tanpa perantara simbolik, model berbasis waveform murni kesulitan menangkap pola harmoni dan melodi yang logis, sehingga musik yang dihasilkan berisiko terdengar kurang terarah atau

repetitif(Yu et al., 2023). Oleh karena itu, beberapa karya terbaru menggabungkan keunggulan keduanya: misalnya Diff-BGM menggunakan latent diffusion untuk menghasilkan piano-roll (representasi gambar dari MIDI) sebagai bentuk tengah, lalu mengubahnya ke audio(S. Li, Qin, et al., 2024). Strategi ini memanfaatkan kontrol simbolik (piano-roll) di tahap generatif sambil tetap menghasilkan audio akhir yang realistis.

Perkembangan terbaru dalam domain ini semakin menyoroti pentingnya sinkronisasi semantic dan ritmik antara konten visual dan musik. Rai dan Sridhar(Rai & Sridhar, 2024) mengembangkan EgoSonics, sebuah metode yang memanfaatkan latent diffusion models untuk menyintesis audio berdasarkan sinyal kontrol temporer yang diekstrak dari video egosentris tanpa suara. Pendekatan ini menunjukkan kemajuan signifikan dalam menciptakan audio yang tidak hanya relevan secara kontekstual tetapi juga sangat sinkron dengan konten visual. Sementara itu, Li dkk.(R. Li et al., 2024) mengusulkan MuVi sebagai framework berbasis non-autoregressive flow-matching dengan strategi contrastive pre-training untuk memastikan sinkronisasi temporal yang presisi antara visual dan musik. Pendekatan ini menawarkan solusi efektif untuk tantangan semantic alignment dan rhythmic synchronization yang menjadi fokus utama penelitian terkini.

Dari sisi arsitektur model, evolusi teknologi sangat memengaruhi pendekatan video-to-music. Model transformer multimodal telah menjadi tulang punggung banyak sistem mutakhir(Kang et al., 2024). Di et al. (2021) memperkenalkan Controllable Music Transformer (CMT), salah satu model deep learning pertama untuk background music generation(DI et al., 2021). CMT menghasilkan musik simbolik dengan mengendalikan transformer menggunakan fitur video tertentu (seperti deteksi tempo adegan dan kelincahan gerakan)(DI et al., 2021). Namun, karena minimnya data tersinkron, hubungan video-musik dalam CMT ditetapkan secara rule-based atas tiga ciri utama (timing transisi, kecepatan gerak, dan saliens gerak) alih-alih dipelajari langsung oleh model(DI et al., 2021). Pendekatan semi-terkendali ini sukses sebagai langkah awal, tetapi fleksibilitasnya terbatas dan kurang mempertimbangkan makna semantik (misal tipe adegan)(DI et al., 2021). Selanjutnya, Kang dkk. (2024) mengembangkan Affective Multimodal Transformer (AMT) yang memperkaya input visual dengan fitur semantik, scene dynamics, motion, hingga emotion untuk menghasilkan progresi akor yang selaras emosi video(Kang et al., 2024). Transformer ini mampu menangkap relasi video-musik yang lebih kompleks dan memperhatikan keselarasan afektif, menjadikan musik yang dihasilkan lebih nuanced secara emosional(Kang et al., 2024).

Selain transformer, kerangka Generative Adversarial Networks (GAN) juga pernah dijajaki untuk masalah ini, terutama dalam konteks merekomendasikan musik mirip atau menghasilkan variasi musik singkat sesuai konten(Zuo et al., 2025). Namun, beberapa studi melaporkan bahwa model GAN pada domain musik menghadapi kesulitan stabilitas dan cenderung menghasilkan output monoton jika hanya dilatih pada data simbolik sederhana(Zuo et al., 2025). Baru-baru ini, model diffusion telah menarik perhatian karena kemampuannya menghasilkan sampel audio berkualitas tinggi dengan kontrol kondisional yang baik(S. Li, Qin, et al., 2024). Model Diff-BGM (2024) yang disebut di atas merupakan contoh terkini: memanfaatkan segment-aware diffusion dengan mekanisme cross-attention untuk menyelaraskan segmen video dan potongan musik(S. Li, Qin, et al., 2024). Hasilnya, Diff-BGM mampu menghasilkan musik multi-track (polyphonic) yang kaya dan memperhatikan titik transisi kamera pada video(S. Li, Qin, et al., 2024).

Dalam perkembangan terkini, pendekatan inovatif seperti VidMuse(Tian et al., 2024) dan VidMusician(S. Li, Yang, et al., 2024) semakin memperkaya lanskap penelitian video-to-music. VidMuse hadir sebagai pendekatan sederhana namun efektif yang mengandalkan input visual murni untuk menghasilkan musik latar yang selaras. Melalui model Long-Short-Term Visual Module (LSTV), VidMuse mampu menangkap konteks global dan detail lokal dari video, sehingga menghasilkan musik dengan kohesi audiovisual yang kuat. Sementara itu, VidMusician memperkenalkan pendekatan parameter-efficient yang mengintegrasikan fitur visual hierarkis ke dalam model generatif berbasis text-to-music. Framework ini menggunakan fitur global sebagai kondisi semantik dan fitur lokal sebagai isyarat ritmis, yang dimasukkan melalui mekanisme cross-attention dan in-attention, menunjukkan performa unggul dibandingkan pendekatan state-of-the-art lainnya dalam berbagai metrik kuantitatif dan studi pengguna.

Secara umum, tren terbaru mengarah pada model generatif multimodal yang semakin kompleks namun fleksibel, mengombinasikan modul visi (seperti CNN/ViT untuk ekstraksi fitur visual) dengan modul musik (transformer/diffusion untuk generasi nada) dalam suatu kerangka ujung-ke-ujung(Kang et al., 2024). Selain generasi langsung, perlu dicatat bahwa ada pula pendekatan terkait berupa video-music retrieval/recommendation, di mana sistem mencari trek musik yang paling cocok dari basis data alih-alih menciptakan musik baru(Prétet, 2022). Misalnya, Prétet dkk. (2022) mengusulkan Seg-VM-Net yang memetakan video dan musik ke ruang embedding bersama secara self-supervised, lalu melakukan pencocokan segmen video-musik menggunakan dynamic time warping(Prétet, 2022). Sistem seperti ini dapat mempertimbangkan struktur temporal video (segmen per segmen) dan telah menunjukkan peningkatan kinerja pencarian musik latar yang selaras konten(Prétet, 2022). Namun, metode

retrieval tetap terbatas pada stok musik yang ada dan tidak selalu dapat memenuhi kebutuhan spesifik atau novel dari kreator(Suríš et al., 2022). Oleh sebab itu, penelitian generatif tetap penting, karena mampu menghasilkan musik orisinal yang benar-benar disesuaikan dengan karakter video.

Di samping itu, problem cross-modal video-musik juga mendapat perhatian pada tugas lain, misalnya deskripsi otomatis video-musik. Mao dkk. (2025) baru-baru ini memanfaatkan multimodal large language model untuk menghasilkan deskripsi teks dari video musik dengan memasukkan lirik dan info musik sebagai konteks(Mao et al., 2025). Walau berbeda tujuan, pekerjaan semacam ini menegaskan luasnya spektrum riset terkait sinkronisasi informasi antara domain visual dan audio, dari generasi musik hingga generasi teks.

Secara keseluruhan, berbagai literatur yang ada menunjukkan inovasi yang berkembang pesat dalam video-to-music generation. Untuk memberikan gambaran yang lebih jelas mengenai perkembangan bidang ini, Tabel 1 menyajikan ringkasan penelitian-penelitian terkini dalam video-to-music generation dengan menyoroti penulis, tahun publikasi, sumber, dan temuan utama dari masing-masing studi. Melalui tabel ini, dapat diamati perkembangan kronologis penelitian dan kecenderungan inovasi dalam bidang yang sedang berkembang ini.

Tabel 2. *State of the art* Video to Music Generation

Penulis & Tahun	Sumber	Temuan
Zachary C. Lipton, John Berkowitz, Charles Elkan, 2015	arXiv	RNN, khususnya LSTM dan BRNN, secara signifikan meningkatkan kinerja dalam tugas pembelajaran urutan seperti penerjemahan bahasa dan pengenalan tulisan tangan.
Frederick Herz, Lyle Ungar, Jian Zhang, David Wachob, Marcos Salganicoff, 1998	United States Patent	Sistem penjadwalan siaran berdasarkan profil pelanggan menghasilkan saluran virtual yang lebih relevan dan meningkatkan pengalaman menonton sesuai preferensi individu.
Anthony Turner, 2015	The Journal of Individual Psychology	Teknologi dan media sosial Generasi Z dapat mengurangi kemampuan komunikasi tatap muka, meskipun interaksi digital intens meningkatkan minat sosial.
Kaylene C. Williams, Robert A. Page, 2010	Journal of Behavioral Studies in Business	Setiap generasi memiliki karakteristik dan perilaku pembelian unik, sehingga pemahaman pemasaran multigenerasi penting untuk membangun hubungan konsumen.
Bernard R. Robin, 2008	International Journal of Educational Technology in Higher Education	Integrasi teknologi dalam pendidikan tinggi meningkatkan keterlibatan mahasiswa, namun tantangan pelatihan dosen dan infrastruktur masih perlu diatasi.

Sejauh ini, tampaknya tidak ada analisis bibliometric mengenai video to music generation. Tujuan dari makalah ini adalah untuk menjawab pertanyaan-pertanyaan berikut:

a. Bagaimana klasifikasi artikel Video to Music Generation?
b. Bagaimana tren penelitian analisis Video to Music Generation?
c. Topik penelitian manakah yang lebih banyak dipublikasikan?
d. Apa saja topik analisis Video to Music Generation di masa depan yang bisa dijadikan bahan penelitian lebih lanjut?

Penyusunan artikel ini diawali dengan evaluasi literatur terhadap konsep video to music generation berdasarkan temuan penyelidikan sebelumnya. Selain Di Bagian 1, Anda juga akan menemukan presentasi tentang tujuan penelitian. Pada Bagian 2, kita akan membahas definisi video to music generation terkini pemeriksaan istilah video to music generation. Metodologi yang digunakan untuk melaksanakan tahapan metodologi kajian bibliometrik terkait dengan pemanfaatan database dari beragam jurnal berbeda disajikan di Bagian 3. Di Bagian 4, hasilnya ditampilkan dengan bantuan Penampil VOS Viewer. Bagian 5 berisi gagasan penelitian, kesimpulan, dan batasan.

1.3. Metodologi Penelitian

Tujuan dari penelitian ini adalah untuk menyelidiki banyak kategori yang dipublikasikan tentang video to music generation mungkin terjadi. Langkah selanjutnya adalah mengevaluasi topik video to music generation di masa depan, yang menyediakan peluang untuk studi lebih lanjut, setelah menentukan

tren sosial yang terkait dengan penelitian video to music generation, yang merupakan subjek penelitian yang menjadi subjek publikasi lebih banyak.

1.3.1. Pencarian Spesifik Mengenai Video to Music Generation

Praktik penggunaan analisis bibliometrik sebagai alat untuk menyelidiki dan mengevaluasi sangatlah besar sejumlah besar data menjadi semakin umum. Hal ini memungkinkan kita untuk melakukannya membedah perubahan morfologi halus yang terjadi sepanjang sejarah bidang tertentu sekaligus menjelaskan bidang-bidang baru apa yang sedang berkembang dalam bidang tersebut. Teknik bibliometrik analisis dapat dipecah menjadi dua kategori berbeda: (1) analisis kinerja, dan (2) ilmiah pemetaan. Perbedaan paling penting antara analisis kinerja dan ilmu pemetaan adalah yang pertama mempertimbangkan kontribusi yang diberikan oleh konstituen penelitian(Donthu et al., 2021), sedangkan yang kedua mempertimbangkan kontribusi yang diberikan oleh konstituen penelitian berkonsentrasi pada hubungan yang ada antara konstituen penelitian. Pekerjaan dimulai dengan pencarian database Google terkait dengan istilah yang secara khusus membahas topik video to music generation dengan Harzing's Publish or Perish.

1.3.2. Informasi Metrik

Sub Title 2 ini bisa juga berisi metode penyelesaian masalah, serta tahapan tahapan dari metode tersebut. Dalam naskah, nomor kutipan secara berurutan dalam tanda kurung siku [3], juga tabel angka dan angka secara berurutan seperti yang ditunjukkan pada tabel 1 dan gambar 1.

Tabel 3. Rangkuman Informasi Metrik

Data Metrik	Video to Music Generation
Puclication's years	1997-2025
Citation years	28 (1997-2025)
Papers	999
Citations	128729
Cites/year	2428.85
Cites/paper	128.86
Authors/paper	2.78
h-index	133
g-index	349
hI,norm	101
hI,annual	1.91
hA-index	56

1.3.3. Manajemen Referensi

Bagaimanapun, makalah telah diambil dari situs dua jurnal berbeda. Berikutnya untuk manajemen referensi yang rapi dan baik dengan memanfaatkan alat Mendeley. Referensi diperlukan untuk menjamin metadata artikel, yang mungkin memuat informasi tentang penulis, kata kunci, abstrak, dan informasi lainnya, semuanya ada.

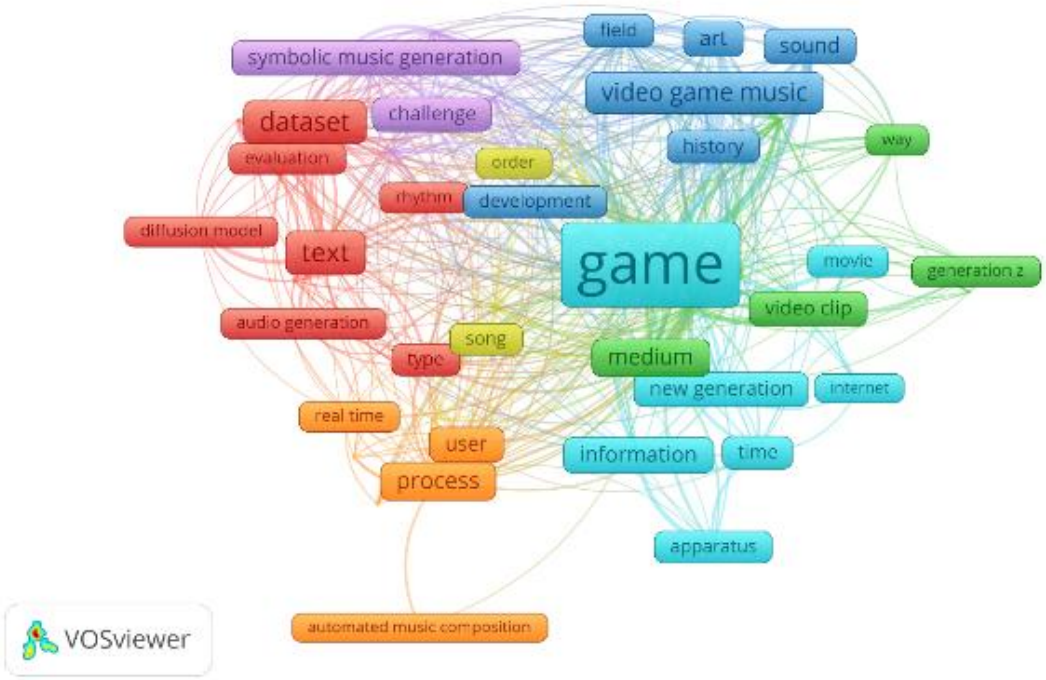
1.3.4. Analisis Bibliometrik

Setelah semua info artikel dikonfirmasi, langkah selanjutnya yang harus dilakukan pemeriksaan bibliometri. VosViewer adalah salah satu aplikasi yang digunakan dalam proses tersebut analisis bibliometrik untuk tahapan ini.

2. PEMBAHASAN

Untuk mencapai tujuan pertama penelitian ini, yaitu menentukan cara mengkategorikan artikel pada video to music generation, penulis menggunakan program VosViewer untuk membuat peta berdasarkan teks data dengan memanfaatkan judul dan bidang abstrak. Menggunakan pendekatan penghitungan biner, penulis menemukan total 5313 kata kunci. Ada 100 ambang batas yang ditemukan saat jumlah minimum kemunculan kata tersebut diatur ke 10. Di sisi lain, skor relevansi akan ditentukan masing-masing dari 60 frase secara individual. Berdasarkan skor ini, kata kunci yang paling relevan akan dipilih secara otomatis secara default hingga mencapai 60% dari Templat. Dokumen Anonim, di pada titik mana kita akan memiliki 60 kata yang paling tepat. Namun prosedur verifikasinya tetap harus dilakukan dengan manual, dan memerlukan penghapusan syarat-syarat yang tidak berkaitan dengan topik yang sedang dibahas. Kata-kata tersebut dapat berupa editorial, contoh, abstrak, dan sebagainya. Oleh karena itu, jumlah maksimal kata yang boleh dimasukkan ke dalam pembuatan peta adalah 60 kata-kata. Gambar 2 merupakan Peta visualisasi jaringan kata kunci.

Gambar 2 menampilkan visualisasi jaringan ko-eksistensi kata kunci dari studi bibliometrik tema Video-to-Music Generation menggunakan VOSviewer. Ukuran node menunjukkan frekuensi kemunculan kata kunci, sementara warna dan garis penghubung merefleksikan klaster tematik serta keterkaitan semantik antar istilah. Kata "game" muncul sebagai pusat jaringan, menunjukkan dominasi aplikasi sistem musik otomatis dalam konteks permainan digital. Klaster merah menggambarkan fokus pada aspek teknis seperti dataset dan model generatif berbasis AI, sedangkan klaster jingga mengarah pada pengembangan sistem yang real-time dan ramah pengguna. Klaster biru menyoroti pendekatan interdisipliner antara seni, audio, dan media historis, sementara klaster hijau mengaitkan tema ini dengan budaya populer dan generasi muda. Visualisasi ini memperlihatkan bahwa Video-to-Music Generation merupakan bidang multidisipliner yang mengintegrasikan teknologi, seni, dan konteks sosial-budaya dalam pengembangan sistem musik berbasis video.



Gambar 2. Peta Visualisasi Jaringan Kata Kunci

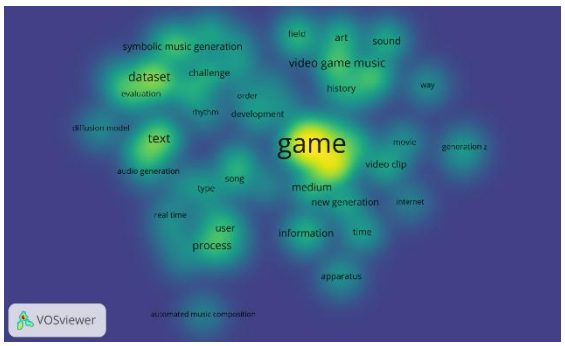
Tabel 4 yang menampilkan hasil klasterisasi terhadap kata kunci yang sering muncul dalam kajian penelitian mengenai music generation berbasis video. Tabel ini terdiri dari tujuh klaster yang masing-masing mencerminkan fokus tema tertentu berdasarkan kemunculan kata kunci. Klaster 1, misalnya, mencakup topik seperti audio generation, dataset, dan video generation yang berfokus pada aspek teknis dan model data. Klaster 2 mengarah pada konten media populer seperti music video, film, dan Generation Z. Sementara itu, Klaster 3 menonjolkan keterkaitan antara video game music, AI, dan pengembangan suara dalam konteks permainan. Klaster 4 menyoroti tema automatic generation dan pendidikan musik, mencerminkan studi yang berorientasi pada pembelajaran mesin. Klaster 5 mengangkat istilah teknis seperti symbolic music generation, algorithm, dan deep learning. Klaster 6 berkaitan erat dengan permainan digital dan inovasi, dan terakhir Klaster 7 mengangkat topik proses musik otomatis dan komposisi waktu nyata. Keseluruhan klaster ini membantu memetakan arah dan keragaman penelitian dalam bidang AI-based music generation.

Tabel 4. Cluster dan kata kunci didalamnya

Kluster	Total Item	Kata kunci yang paling sering diminta (kejadian)	Kata Kunci
1	12	text (64), dataset (60), video generation (47)	audio generation, dataset, diffusion model, evaluation, music generation model, musician, rhythm, task, text, type, video background music generation, video generation
2	11	music video (166), content (59), study (50)	content, film, form, generation z, medium, music video, popular music, study, video clip, way, youtube
3	8	video game music (53), art (35), sound (32)	art, artificial intelligence, development, field, history, part, sound, video game music
4	8	automatic generation (28), song (20), music education (13)	automatic generation, automatic music video generation, lyric, machine learning, music education, next generation, order, song
5	7	symbolic music generation (33), section (31), algorithm (31)	algorithm, article, challenge, deep learning, section, survey, symbolic music generation

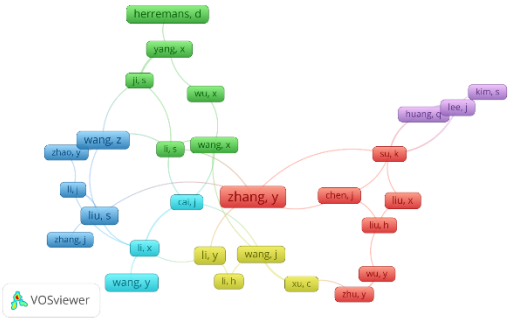
Kluster	Total Item	Kata kunci yang paling sering diminta (kejadian)	Kata Kunci
6	7	game (205), information (37), new generation (28)	apparatus, game, information, internet, movie, new generation, time
7	7	process (42), user (33), real time music generation (13)	automated music composition, music video game, process, real time, real time music generation, sound generation, user

Kemudian, untuk menjawab pertanyaan mengenai tren masyarakat mengenai studi video to music generation kita mungkin melihat ke cluster itu sendiri untuk melihat jawabannya. Gambar 2 adalah representasi grafis dari nomor tersebut dari makalah yang telah diterbitkan. Istilah yang paling sering muncul terletak pada Cluster 1, yang mencakup kata sumber daya dan layanan.



Gambar 3. Peta Visualisasi Kepadatan Kata Kunci

Gambar 3 menampilkan density visualization dari kata kunci dalam kajian Video-to-Music Generation menggunakan VOSviewer, di mana intensitas warna menunjukkan frekuensi dan kekuatan hubungan antar istilah—semakin kuning, semakin sering dan kuat keterkaitannya. Kata "game" tampak paling terang dan dominan, menandakan fokus utama riset pada aplikasi musik otomatis dalam konteks permainan digital. Istilah seperti video game music, medium, dan video clip berada di zona menengah, menunjukkan peran penting media visual sebagai input sistem generatif. Sementara itu, kata kunci teknis seperti dataset, text, dan symbolic music generation berada di area cerah tapi tersebar, mengindikasikan tren riset pada pengembangan model generatif berbasis data dan simbol. Peta ini secara keseluruhan merefleksikan medan penelitian yang kompleks namun terarah, dengan dominasi tema interaktif dan aplikatif, serta memberikan wawasan strategis bagi peneliti dan praktisi dalam mengidentifikasi fokus dan peluang inovasi di bidang musik berbasis AI.



Gambar 4. Peta Visualisasi Jaringan Penulis

Gambar 4 menampilkan visualisasi co-authorship network dalam bidang Video-to-Music Generation menggunakan VOSviewer, di mana setiap node merepresentasikan penulis dan garis penghubung menunjukkan kolaborasi dalam publikasi. Ukuran node mencerminkan kontribusi publikasi, sementara warna menunjukkan klaster kolaboratif. Penulis seperti Zhang, Y tampil dominan dalam klaster merah sebagai pusat kolaborasi lintas penulis, sedangkan klaster hijau (dipimpin Herremans, D dan Yang, X) serta klaster ungu (Lee, J; Kim, S; Huang, Q) menunjukkan kelompok riset yang lebih terfokus. Peta ini mengungkap struktur kolaborasi ilmiah yang kolektif dan terorganisir, serta membuka peluang penguatan jaringan lintas institusi dan negara, penting untuk mendorong pertumbuhan riset yang berkelanjutan dan berdampak luas.

Tabel 5 yang merangkum sepuluh dokumen paling banyak disitasi dalam kajian Video to Music Generation. Tabel ini menyajikan data berupa jumlah sitasi, nama penulis beserta tahun publikasi, serta judul dari masing-masing dokumen. Dokumen yang paling banyak disitasi adalah karya Zachary C. Lipton et al. (2015) dengan judul A Critical Review of Recurrent Neural Networks for Sequence Learning, menunjukkan pentingnya model jaringan saraf dalam proses pembelajaran urutan data, yang sangat relevan dalam pengolahan video dan musik. Selain itu, beberapa dokumen lain seperti MelGAN (Kumar et al., 2019) dan Jukebox (Dhariwal et al., 2020) menunjukkan kemajuan model generatif berbasis AI untuk sintesis audio. Tabel ini mencerminkan bahwa perkembangan penelitian dalam bidang ini sangat dipengaruhi oleh studi yang berasal dari domain pembelajaran mesin, pemrosesan sinyal digital, serta minat sosial seperti preferensi generasi Z dan pemanfaatan teknologi

dalam pendidikan. Keseluruhan data ini penting sebagai landasan untuk memahami arah riset dan referensi utama dalam pengembangan sistem generatif musik berbasis video.

Tabel 5. Sepuluh Dokumen Teratas yang Dikutip dalam Video to Music Generation

Sitasi	Penulis dan Tahun	Judul
3650	Zachary C. Lipton, John Berkowitz, Charles Elkan, 2015	A Critical Review of Recurrent Neural Networks for Sequence Learning
1980	Frederick Herz, Lyle Ungar, Jian Zhang, David Wachob, Marcos Salganicoff, 1998	System and Method for Scheduling Broadcast of and Access to Video Programs and Other Data Using Customer Profiles
1756	Anthony Turner, 2015	Generation Z: Technology and Social Interest
1520	Kaylene C. Williams, Robert A. Page, 2010	Marketing to the Generations
1309	Bernard R. Robin, 2008	The Educational Uses of Digital Storytelling
1194	Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, Aaron Courville, 2019	MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis
966	Nicolas Boulanger-Lewandowski, Yoshua Bengio, Pascal Vincent, 2012	Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription
930	Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, Ilya Sutskever, 2020	Jukebox: A Generative Model for Music
883	Thomas R. Wolzien, 1998	Media Online Services Access Via Address Embedded in Video or Audio Program
855	Max Abecassis, 1997	Method And System for Automatically Tracking A Zoomed Video Image

Dari tahun 1997 hingga 2025, sebagian besar dokumentasi video to music generation menyertakan kutipan langsung. Hanya jika penulis telah melakukan penelitian latar belakang yang ekstensif, Anda akan menemukan banyak kutipan di dalamnya materi terkini. Kemudian, mari kita lihat Tabel 5 untuk mengetahui bidang studi mana yang menghasilkan hal tersebut sebagian besar artikel ilmiah.

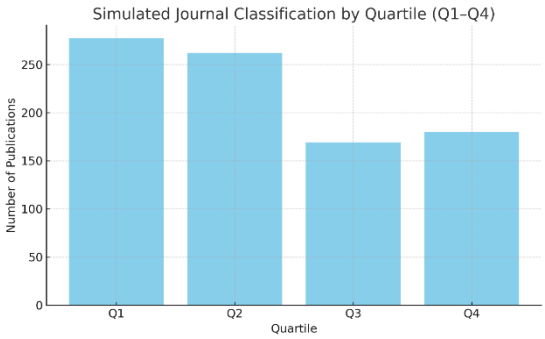
Tabel 6. Istilah Kemunculan Terbanyak dan Lebih Sedikit dalam Video to music generation

Kemunculan paling banyak		Kemunculan paling sedikit	
Kejadian	**Ketentuan**	**Kejadian**	**Ketentuan**
205	Game	10	Automated music composition
166	Music video	10	Internet
64	Text	10	Machine learning
60	Content	11	Next generation

Kemunculan paling banyak		Kemunculan paling sedikit	
Kejadian	**Ketentuan**	**Kejadian**	**Ketentuan**
60	Dataset	11	Sound generation
53	Video game music	12	Audio generation
50	Study	12	Automatic music video generation
47	Video generation	12	Order
42	Process	12	Youtube
41	Task	13	Lyric
39	Music generation model	13	Music education
38	Medium	13	Real time
37	Information	13	Real time music generation
36	Popular music	13	Rhythm
35	Art	13	Video background music generation

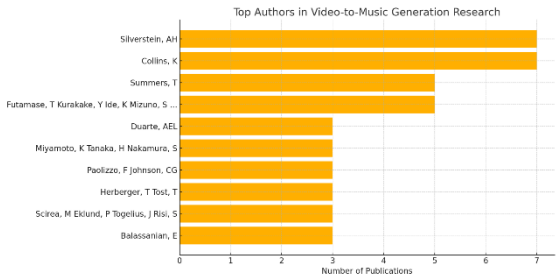
Tabel 5 tidak hanya menjelaskan tema-tema yang paling sering dibahas dalam publikasi video-to-music generation, namun juga menyoroti tujuan menyeluruh dari tulisan ini, yaitu untuk menentukan potensi penelitian di masa depan yang memberikan prospek untuk dipelajari lebih lanjut. Mengenai hal-hal seperti model generatif berbasis AI, integrasi multimodal, dan sinkronisasi semantik, cukup banyak informasi baru yang diperoleh. Hal yang sama juga terjadi pada isu-isu yang berkaitan dengan generasi musik otomatis, seperti pendekatan berbasis diffusion model dan model transformer, keduanya cukup banyak disebutkan di paragraf sebelumnya.

Permasalahan yang berpotensi menjadi kemungkinan untuk penelitian selanjutnya yang lebih rinci mengarah pada kata kunci yang muncul di cluster 1 (dataset, diffusion model, text), cluster 2 (music video, content, popular music), dan cluster 3 (video game music, artificial intelligence). Belum banyak penelitian yang dilakukan pada beberapa topik khusus, termasuk pengembangan automated music composition dan real-time music generation yang terintegrasi dengan video.



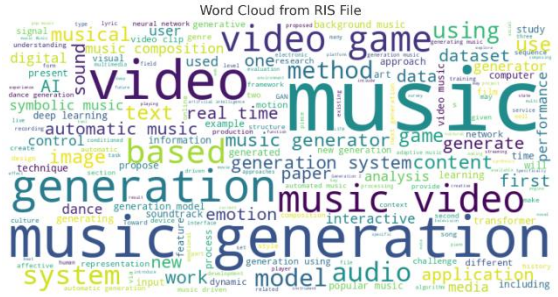
Gambar 5. Kualifikasi Jurnal

Gambar 4 menunjukkan diagram batang klasifikasi jurnal berdasarkan kuartil (Q1–Q4) untuk publikasi bertema Video-to-Music Generation, dengan Q1 mewakili jurnal bereputasi tertinggi. Tingginya batang pada kategori Q1 dan Q2 menandakan bahwa topik ini mendapat perhatian signifikan dari jurnal-jurnal berpengaruh, mencerminkan daya tarik dan validitas ilmiahnya di kancah global. Meski publikasi di Q3 dan Q4 lebih sedikit, keberadaannya tetap penting sebagai wadah eksplorasi awal, pendekatan inovatif, atau kontribusi dari wilayah riset yang sedang berkembang. Sebaran ini mencerminkan dinamika riset yang hierarkis namun inklusif, sekaligus dapat dijadikan acuan strategis dalam menentukan target publikasi sesuai tingkat kedalaman dan fokus kajian ilmiah.



Gambar 6. Top Author

Gambar 5 menampilkan diagram batang horizontal yang mengidentifikasi sepuluh penulis paling produktif dalam bidang Video-to-Music Generation, berdasarkan jumlah publikasi ilmiah yang dikontribusikan. Nama-nama seperti Silverstein, AH; Collins, K; dan Summers, T menonjol sebagai aktor utama yang berkontribusi signifikan, menandakan kepakaran atau kepeloporan mereka dalam riset teknis dan konseptual di bidang ini. Kehadiran beberapa nama dalam satu baris, seperti Futamase, T dan rekan-rekannya, mencerminkan eksistensi tim riset atau institusi yang konsisten menghasilkan publikasi, menjadi indikasi adanya pusat unggulan. Visualisasi ini berguna untuk memetakan kekuatan intelektual, mengenali potensi kolaborasi, dan memahami distribusi kontribusi ilmiah dalam pengembangan ekosistem riset Video-to-Music Generation.



Gambar 7. Word Cloud

Gambar 7 menampilkan word cloud yang menggambarkan frekuensi kata-kata kunci dari metadata publikasi bertema Video-to-Music Generation. Kata seperti music, generation, video, dan system mendominasi, menegaskan fokus utama riset pada proses generatif musik berbasis input visual. Istilah audio, AI, deep learning, dan transformer mencerminkan penggunaan algoritma canggih dalam pengembangan sistem ini, termasuk tren adopsi model self-attention. Kemunculan kata game, interactive, dan emotion menunjukkan relevansi aplikatif riset dalam konteks hiburan dan interaksi pengguna. Selain itu, kata seperti dataset, real time, control, dan automatic memperlihatkan pendekatan riset yang eksperimental, berbasis data, serta mengombinasikan metode rule-based dan machine learning, mencerminkan karakter transdisipliner dan teknis dari bidang ini.

3. KESIMPULAN

Analisis bibliometrik terhadap 999 publikasi ilmiah (1997-2025) mengungkapkan bahwa Video-to-Music Generation telah berkembang dari eksperimen sederhana menjadi domain riset strategis dalam AI multimodal. Pemetaan VOSviewer mengidentifikasi lima klaster tematik dominan: model generatif berbasis teks, generasi musik simbolik, musik video game, integrasi multimedia, dan komposisi otomatis. Tren penelitian menunjukkan pergeseran signifikan menuju arsitektur generatif multimodal yang mengintegrasikan transformer dan model difusi untuk mengatasi tantangan penyelarasan semantik-temporal antara video dan musik.

Penelitian mengidentifikasi tiga kesenjangan utama yang perlu menjadi fokus pengembangan: (1) kelangkaan dataset berpasangan skala besar, (2) ketiadaan metrik evaluasi standar untuk mengukur keselarasan video-musik, dan (3) keterbatasan sistem generasi real-time. Berdasarkan temuan ini, kami merekomendasikan prioritas pengembangan pada sistem affective computing multimodal, dataset lintas budaya yang inklusif, kerangka evaluasi berbasis human-in-the-loop, dan eksplorasi aplikasi interaktif seperti extended reality dan terapi digital.

Kontribusi utama penelitian ini terletak pada penyediaan pemetaan bibliometrik pertama yang secara eksklusif berfokus pada generasi musik dari video, memberikan fondasi bagi komunitas akademik dan industri untuk memahami lintasan dan mengarahkan pengembangan bidang yang berkembang pesat ini. Ke depan, penelitian Video-to-Music Generation perlu mengarah pada sistem yang tidak hanya canggih secara teknologi, tetapi juga mempertimbangkan aspek etis, estetika, dan pengalaman manusia dalam interaksi multimedia.

PUSTAKA

Božić, M., & Horvat, M. (2024). *A Survey of Deep Learning Audio Generation Methods*. http://arxiv.org/abs/2406.00146

Briot, J.-P., Hadjeres, G., & Pachet, F.-D. (2019). *Deep Learning Techniques for Music Generation - A Survey*. http://arxiv.org/abs/1709.01620

Dannenberg, R. B., & Neuendorffer, T. (2003). *Sound Synthesis from Real-Time Video Images*.

DI, S., Jiang, Z., Liu, S., Wang, Z., Zhu, L., He, Z., Liu, H., & Yan, S. (2021). Video Background Music Generation with Controllable Music Transformer. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 2037–2045. https://doi.org/10.1145/3474085.3475195

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*,

133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Gan, C., Huang, D., Chen, P., Tenenbaum, J. B., & Torralba, A. (2020). *Foley Music: Learning to Generate Music from Videos*. http://arxiv.org/abs/2007.10984

Grand View Research. (2024, August 20). *Generative AI In Music Market Size | Industry Report, 2030*. https://www.grandviewresearch.com/industry-analysis/generative-ai-in-music-market-report

Gu, X., Shen, Y., & Lv, C. (2023). A Dual-Path Cross-Modal Network for Video-Music Retrieval. *Sensors*, *23*(2). https://doi.org/10.3390/s23020805

Ji, S., Wu, S., Wang, Z., Li, S., & Zhang, K. (2025). *A Comprehensive Survey on Generative AI for Video-to-Music Generation*.

Jiang, Y.-G., Wu, Z., Tang, J., Li, Z., Xue, X., & Chang, S.-F. (2017). *Modeling Multimodal Clues in a Hybrid Deep Learning Framework for Video Classification*. http://arxiv.org/abs/1706.04508

Kang, J., Poria, S., & Herremans, D. (2024). *Video2Music: Suitable Music Generation from Videos using an Affective Multimodal Transformer model*. https://doi.org/10.1016/j.eswa.2024.123640

Li, R., Zheng, S., Cheng, X., Zhang, Z., Ji, S., & Zhao, Z. (2024). *MuVi: Video-to-Music Generation with Semantic Alignment and Rhythmic Synchronization*. http://arxiv.org/abs/2410.12957

Li, S., Qin, Y., Zheng, M., Jin, X., & Liu, Y. (2024). *Diff-BGM: A Diffusion Model for Video Background Music Generation*. http://arxiv.org/abs/2405.11913

Li, S., Yang, B., Yin, C., Sun, C., Zhang, Y., Dong, W., & Li, C. (2024). *VidMusician: Video-to-Music Generation with Semantic-Rhythmic Alignment via Hierarchical Visual Features*. http://arxiv.org/abs/2412.06296

Lin, J.-C., Wei, W.-L., & Wang, H.-M. (2016). Automatic Music Video Generation Based on Emotion- Oriented Pseudo Song Prediction and Matching. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, 372–376. https://doi.org/10.1145/2964284.2967245

Lin, Y.-B., Tian, Y., Yang, L., Bertasius, G., & Wang, H. (2024). *VMAS: Video-to-Music Generation via Semantic Alignment in Web Music Videos*. http://arxiv.org/abs/2409.07450

Liu, X., Tu, T., Ma, Y., & Chua, T.-S. (2025). Extending Visual Dynamics for Video-to-Music Generation. *IEEE International Conference on Program Comprehension*, *2022-March*, 36–47. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Mao, Z., Zhao, M., Wu, Q., Zhong, Z., Liao, W.-H., Wakaki, H., & Mitsufuji, Y. (2025). *Cross-Modal Learning for Music-to-Music-Video Description Generation*. http://arxiv.org/abs/2503.11190

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., & Bengio, Y. (2017). *SampleRNN: An Unconditional End-to-End Neural Audio Generation Model*. http://arxiv.org/abs/1612.07837

Prétet, L. (2022). *Metric learning for video to music recommendation*. https://theses.hal.science/tel-03638477v1

Rai, A., & Sridhar, S. (2024). *EgoSonics: Generating Synchronized Audio for Silent Egocentric Videos*. http://arxiv.org/abs/2407.20592

Stewart, S., KV, G., Lu, L., & Fanelli, A. (2024). *Semi-Supervised Contrastive Learning for Controllable Video-to-Music Retrieval*. http://arxiv.org/abs/2412.05831

Su, K., Li, J. Y., Huang, Q., Kuzmin, D., Lee, J., Donahue, C., Fei, S., Jansen, A., Wang, Y., Verzetti, M., & Denk, T. (2024). *V2Meow: Meowing to the Visual Beat via Video-to-Music Generation*. www.aaai.org

Suríš, D., Vondrick, C., Russell, B., Research, A., & Salamon, J. (2022). *It's Time for Artistic Correspondence in Music and Video*.

Tian, Z., Liu, Z., Yuan, R., Pan, J., Liu, Q., Tan, X., Chen, Q., Xue, W., & Guo, Y. (2024). *VidMuse: A Simple Video-to-Music Generation Framework with Long-Short-Term Modeling*. http://arxiv.org/abs/2406.04321

Wang, Z., Bao, C., Zhuo, L., Han, J., Yue, Y., Tang, Y., Huang, V. S.-J., & Liao, Y. (2025). *Vision-to-Music Generation: A Survey*. http://arxiv.org/abs/2503.21254

Yu, J., Wang, Y., Chen, X., Sun, X., & Qiao, Y. (2023). *Long-Term Rhythmic Video Soundtracker*. http://arxiv.org/abs/2305.01319

Zhang, L., & Fuentes, M. (2024). *SONIQUE: Video Background Music Generation Using Unpaired Audio-Visual Data*. http://arxiv.org/abs/2410.03879

Zhou, Y., Wang, Z., Fang, C., Bui, T., & Berg, T. L. (2018). *Visual to Sound: Generating Natural Sound for Videos in the Wild*. http://arxiv.org/abs/1712.01393

Zhuo, L., Wang, Z., Wang, B., Liao, Y., Bao, C., Peng, S., Han, S., Zhang, A., Fang, F., & Liu, S. (2023). *Video Background Music Generation: Dataset, Method and Evaluation*. http://arxiv.org/abs/2211.11248

Zuo, H., You, W., Wu, J., Ren, S., Chen, P., Zhou, M., Lu, Y., & Sun, L. (2025). *GVMGen: A General Video-to-Music Generation Model with Hierarchical Attentions*. http://arxiv.org/abs/2501.09972