BAB II
KAJIAN PUSTAKA

2.1 Dasar Teori

Bagian ini menguraikan konsep teoretis yang menjadi landasan penelitian. Pembahasan mencakup paradigma sistem rekomendasi, pembelajaran multimodal, model representasi berbasis Transformer, mekanisme fusi fitur, pemodelan graf kolaboratif, masalah *cold-start*, teknik evaluasi, serta *dataset benchmark* untuk membangun sistem rekomendasi dengan representasi item yang informatif.

a. Sistem Rekomendasi dan Paradigma Utamanya

Sistem rekomendasi pada platform *streaming* modern seperti Netflix dirancang untuk memahami selera pengguna secara personal dan menyajikan konten yang paling relevan secara otomatis (Steck et al., 2021). Secara fungsional, alur kerjanya terbagi dalam tiga tahapan: pertama, sistem menerima sinyal input dari pengguna berupa judul film yang dicari, film yang sedang ditonton, atau riwayat penilaian yang pernah diberikan; kedua, sistem memproses sinyal tersebut menggunakan algoritma untuk menghitung tingkat kemiripan atau relevansi antara item referensi dengan seluruh katalog yang tersedia; ketiga, sistem menyajikan sejumlah item teratas yang diurutkan berdasarkan skor relevansi sebagai output rekomendasi (Ko et al., 2022). Sebagai contoh konkret, ketika pengguna mengetikkan "Doraemon" di kolom pencarian, sistem tidak sekadar mencocokkan nama secara tekstual, melainkan menganalisis atribut film tersebut seperti genre animasi, tema persahabatan, target usia, hingga karakter yang dominan dalam narasi, sehingga dapat merekomendasikan judul seperti "Ninja Hattori" atau "Crayon Shinchan" yang memiliki kesamaan konten signifikan (Zhang et al., 2021).

Content-based filtering adalah pendekatan rekomendasi yang mengandalkan kemiripan fitur antar item. Sistem membangun representasi setiap film berdasarkan atributnya seperti genre, sinopsis, informasi pemeran, dan fitur visual poster, kemudian mengukur kedekatan antara item referensi dengan item lain menggunakan teknik seperti *cosine similarity*. Kelebihan utama pendekatan ini adalah kemampuannya merekomendasikan item baru yang belum memiliki riwayat

interaksi, karena keputusan sepenuhnya didasarkan pada karakteristik konten item itu sendiri (Deldjoo et al., 2021). Namun kelemahannya, sistem cenderung menghasilkan rekomendasi yang monoton dan terlalu sempit karena hanya menyarankan item yang sangat mirip dengan preferensi pengguna sebelumnya, tanpa ruang untuk penemuan konten baru yang masih relevan secara selera (Wu et al., 2021).

Collaborative filtering, sebaliknya, tidak bergantung pada atribut konten sama sekali, melainkan berangkat dari asumsi bahwa pengguna dengan pola interaksi serupa cenderung memiliki selera yang mirip. Jika pengguna A dan B sama-sama menyukai film X dan Y, maka film Z yang disukai A kemungkinan besar juga relevan bagi B, meskipun Z tidak memiliki kemiripan konten yang jelas dengan X maupun Y. Kekuatan pendekatan ini terletak pada kemampuannya menangkap preferensi yang bersifat laten dan tidak selalu dapat dijelaskan melalui deskripsi konten (Steck et al., 2021). Namun, keterbatasannya yang paling mendasar adalah masalah *cold-start*, yaitu kondisi di mana sistem tidak mampu memberikan rekomendasi yang memadai untuk pengguna baru atau item baru yang belum memiliki cukup riwayat interaksi, serta masalah *sparsity* ketika matriks interaksi pengguna-item sangat jarang terisi (Rodríguez-Baena et al., 2025).

a. Pendekatan Hibrida dalam Sistem Rekomendasi

Kedua paradigma di atas, meskipun telah terbukti efektif pada konteks tertentu, memiliki keterbatasan yang bersifat saling melengkapi. *Content-based filtering* unggul dalam menangani item baru namun lemah dalam menangkap preferensi kolektif yang laten, sementara *collaborative filtering* efektif dalam memodelkan pola preferensi komunitas namun rentan terhadap kondisi *cold-start* dan *sparsity* data. Kesadaran akan keterbatasan ini mendorong berkembangnya pendekatan *hybrid* yang menggabungkan sinyal dari kedua paradigma secara bersamaan untuk menghasilkan sistem yang lebih *robust* dan akurat (Zhang et al., 2021). Dalam praktiknya, pendekatan *hybrid* terbukti mampu meningkatkan kualitas rekomendasi secara konsisten dibandingkan penggunaan salah satu metode secara terpisah, sebagaimana ditunjukkan dalam berbagai studi komparatif pada *dataset benchmark* seperti MovieLens (Ko et al., 2022).

Alasan mendasar digunakannya pendekatan *hybrid* dalam penelitian ini berkaitan langsung dengan karakteristik data yang dihadapi. *Dataset* MovieLens memiliki tingkat *sparsity* yang sangat tinggi, di mana lebih dari 93% kemungkinan interaksi pengguna-item tidak pernah terjadi. Pada kondisi ini, *collaborative filtering* murni kesulitan membentuk representasi yang informatif, terutama untuk item dengan jumlah interaksi terbatas. Di sisi lain, setiap film pada dasarnya menyimpan informasi yang kaya dari berbagai sumber, mulai dari visual poster, narasi sinopsis, hingga metadata terstruktur seperti informasi pemeran dan genre. Memanfaatkan kekayaan informasi konten ini sebagai pelengkap sinyal kolaboratif merupakan strategi yang secara teoritis dan empiris terbukti efektif dalam mengurangi dampak *sparsity* dan *cold-start* (Deldjoo et al., 2021; Wu et al., 2021).

Lebih jauh, perkembangan arsitektur *deep learning* modern membuka peluang bagi pendekatan *hybrid* yang lebih canggih, di mana pemrosesan konten multimodal dan pemodelan struktur kolaboratif tidak lagi dilakukan secara terpisah, melainkan dalam satu kerangka terintegrasi secara *end-to-end*. (Steck et al., 2021) menunjukkan bahwa sistem rekomendasi skala industri seperti Netflix secara konsisten mengandalkan kombinasi berbagai model yang saling melengkapi, dengan tujuan menyeimbangkan akurasi prediksi dan keberagaman rekomendasi. Penelitian ini mengadopsi filosofi yang sama dengan mengintegrasikan representasi konten multimodal berbasis Transformer sebagai fondasi *embedding* item pada arsitektur graf kolaboratif LightGCN, sehingga informasi konten yang kaya dari berbagai modalitas dapat langsung memperkaya proses pembelajaran pola preferensi kolektif sejak tahap awal propagasi graf (Kim et al., 2023; Xia et al., 2024).

b. Multimodal Learning dalam Sistem Rekomendasi

Multimodal learning adalah pendekatan pembelajaran mesin yang menggunakan lebih dari satu sumber informasi untuk membentuk representasi yang lebih lengkap. Dalam sistem rekomendasi, berbagai sinyal seperti konten visual, metadata, dan informasi kontekstual dapat digabungkan untuk memperkaya representasi item dan pengguna (Deldjoo et al., 2021).

Integrasi multimodal memungkinkan model menangkap karakteristik item secara lebih menyeluruh. Representasi visual dari poster film memberikan informasi tentang estetika dan genre, representasi tekstual dari sinopsis menangkap plot dan tema naratif, sedangkan metadata terstruktur seperti informasi *cast* dan genre mencerminkan konteks produksi dan kategorisasi film. Kombinasi berbagai modalitas ini meningkatkan akurasi rekomendasi serta ketahanan terhadap kompleksitas dan keragaman data (Liu et al., 2025; Mu & Wu, 2023).

c. Model Transformer untuk Representasi Fitur

Transformer adalah arsitektur berbasis mekanisme *attention* yang mampu memodelkan dependensi global dalam data secara efektif. Dalam sistem rekomendasi, Transformer digunakan sebagai model representasi untuk mengekstraksi fitur tingkat tinggi dari berbagai jenis data karena kemampuannya menangkap hubungan kompleks antar elemen *input* tanpa bergantung pada struktur lokal (Sun et al., 2019).

Vision Transformer (ViT) telah terbukti efektif dalam mengekstraksi fitur visual dari gambar dengan membagi gambar menjadi *patches* dan memodelkan hubungan global antar *patches* tersebut melalui mekanisme *self-attention* (Palanisamy et al., 2025). Pendekatan ini memungkinkan penangkapan informasi visual yang kaya dari poster film, termasuk komposisi, warna, dan elemen visual lainnya yang mencerminkan genre dan mood film (Álvarez-Sánchez et al., 2025).

BERT (*Bidirectional Encoder Representations from Transformers*) memungkinkan pemahaman konteks tekstual yang mendalam dari sinopsis film melalui mekanisme *pre-training* pada data teks berskala besar (Devlin et al., n.d.). Berbeda dengan model sekuensial tradisional, BERT memodelkan konteks secara bidirectional sehingga dapat menangkap nuansa semantik yang kompleks dalam deskripsi naratif film.

Kemampuan arsitektur Transformer dalam memodelkan dependensi global dari data terstruktur maupun tidak terstruktur telah terbukti andal dalam memproses hubungan spasial yang kompleks. Riset terdahulu menunjukkan bahwa model berbasis Transformer efektif dalam menangkap ekstraksi fitur berdimensi tinggi dan variasi spasial yang dinamis, seperti yang diimplementasikan pada analisis pose

pemain olahraga (Haq et al., 2025). Karakteristik penangkapan hubungan relasional jarak jauh (*long-range dependencies*) tanpa kehilangan informasi spasial lokal tersebut melandasi justifikasi penggunaan Vision Transformer (ViT) dalam penelitian ini untuk memproses karakteristik visual poster film secara *end-to-end*.

d. Mekanisme Fusi Multimodal Berbasis Attention

Fusi multimodal berbasis *attention* memungkinkan integrasi berbagai representasi fitur dengan mempelajari bobot kontribusi setiap sumber informasi secara adaptif. Mekanisme *cross-attention* secara khusus memfasilitasi pembelajaran korespondensi antar modalitas dengan memungkinkan setiap modalitas untuk "mengamati" informasi dari modalitas lain. Melalui proses ini, model dapat menyeimbangkan peran masing-masing sinyal sehingga terbentuk representasi item terintegrasi yang lebih informatif dan relevan untuk proses rekomendasi (Kim et al., 2023).

Secara matematis, mekanisme *cross-attention* antara modalitas visual $v$dan tekstual $t$dapat direpresentasikan melalui persamaan (2.1):

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)\text{V} \qquad (2.1)$$

di mana $\text{Q} = \text{vW}_Q$, $\text{K} = \text{tW}_K$, dan $\text{V} = \text{tW}_V$ adalah proyeksi *query*, *key*, dan *value*, dengan d_k adalah dimensi dari *key*. Persamaan (2.1) ini memungkinkan modalitas visual untuk "bertanya" kepada modalitas tekstual tentang informasi relevan, dan sebaliknya, sehingga tercipta representasi yang saling melengkapi (Xia et al., 2024).

e. Graph Neural Network dan LightGCN dalam Sistem Rekomendasi

Graph Neural Network (GNN) digunakan untuk memodelkan hubungan antar entitas dalam sistem rekomendasi melalui struktur graf bipartit yang menghubungkan pengguna dan item. LightGCN adalah varian GNN yang menyederhanakan proses pembelajaran graf dengan fokus pada agregasi linear

neighbor dan menghilangkan transformasi fitur serta fungsi aktivasi nonlinear (X. He et al., 2020). Pendekatan ini menghasilkan model yang efisien secara komputasi dan stabil dalam pembelajaran, sehingga cocok digunakan sebagai komponen *collaborative filtering* yang dapat dikombinasikan dengan representasi fitur eksternal (LeeGeon et al., 2025).

Propagasi sinyal pada LightGCN didefinisikan melalui persamaan (2.2) dan persamaan (2.3):

$$e_u^{(k+1)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} e_i^{(k)} \tag{2.2}$$

$$e_i^{(k+1)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} e_u^{(k)} \tag{2.3}$$

di mana $e_u^{(k)}$ dan $e_i^{(k)}$ adalah *embedding* pengguna u dan item i pada *layer* ke-k, $\mathcal{N}_u$ dan $\mathcal{N}_i$ adalah himpunan tetangga pada graf bipartit. Normalisasi simetris dalam persamaan (2.2) dan (2.3) memastikan stabilitas numerik selama propagasi sinyal (X. He et al., 2020) *Embedding* akhir diperoleh melalui agregasi dari semua *layer* persamaan (2.4) :

$$e_u = \frac{1}{K+1} \sum_{k=0}^{K} e_u^{(k)} \tag{2.4}$$

f. *Cold-start* Problem dan Sparsity dalam Sistem Rekomendasi

Cold-start problem adalah tantangan fundamental dalam sistem rekomendasi yang terjadi ketika sistem harus memberikan rekomendasi untuk item baru atau pengguna baru yang memiliki sedikit atau bahkan tidak ada data interaksi historis. Masalah ini dapat dikategorikan menjadi dua jenis: *item cold-start* yang terjadi ketika item baru ditambahkan ke katalog tanpa riwayat interaksi, dan *user*

cold-start yang terjadi ketika pengguna baru bergabung tanpa riwayat preferensi (Z. Zhou et al., 2023).

Sparsity mengacu pada kondisi di mana matriks interaksi pengguna-item sangat jarang terisi, dengan sebagian besar entri kosong karena setiap pengguna hanya berinteraksi dengan sebagian kecil dari total item yang tersedia. Pada *dataset* MovieLens, tingkat *sparsity* dapat mencapai lebih dari 99%, yang berarti kurang dari 1% dari semua kemungkinan interaksi pengguna-item benar-benar terjadi (Rodríguez-Baena et al., 2025).

Pendekatan multimodal menawarkan solusi untuk masalah *cold-start* dengan memanfaatkan informasi konten item seperti poster, sinopsis, dan metadata. Bahkan tanpa data interaksi, sistem dapat membentuk representasi item yang informatif berdasarkan konten visual dan tekstual, sehingga tetap dapat memberikan rekomendasi yang relevan.

g. *Implicit feedback* dan Bayesian Personalized Ranking

Sistem rekomendasi dapat menggunakan dua jenis *feedback*: *explicit feedback* berupa rating numerik yang diberikan pengguna secara eksplisit, dan *implicit feedback* berupa perilaku pengguna seperti klik, view, atau pembelian yang tidak secara langsung mengindikasikan preferensi (Herlocker et al., 2004). Meskipun *explicit feedback* memberikan sinyal preferensi yang jelas, data *implicit feedback* lebih mudah dikumpulkan dan tersedia dalam jumlah yang jauh lebih besar.

Dalam konteks penelitian ini, data rating dari MovieLens dikonversi menjadi *implicit feedback* dengan memperlakukan rating sebagai interaksi positif (observed) dan tidak adanya rating sebagai interaksi negatif (unobserved). Pendekatan ini memungkinkan penggunaan teknik pembelajaran berbasis ranking yang lebih sesuai untuk skenario rekomendasi praktis.

Bayesian Personalized Ranking (BPR) adalah *loss function* yang dirancang khusus untuk pembelajaran dari *implicit feedback* dengan paradigma pairwise ranking (Rendle et al., 2009) BPR mengoptimalkan model untuk memprediksi bahwa item yang berinteraksi dengan pengguna (positif) harus memiliki skor lebih

tinggi dibandingkan item yang tidak berinteraksi (negatif). Fungsi objektif BPR didefinisikan dalam persamaan (2.5) :

$$\mathcal{L}_{BPR} = - \sum_{(u,i,j) \in D_S} \ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}) + \lambda \parallel \Theta \parallel^2 \quad (2.5)$$

di mana $\hat{y}_{ui}$ dan $\hat{y}_{uj}$ adalah skor prediksi untuk item positif $i$dan item negatif j bagi pengguna u, σ adalah fungsi sigmoid, D_S adalah korpus triplet (u, i, j) di mana pengguna u lebih menyukai item i daripada item j, dan λ adalah parameter regularisasi untuk mencegah *overfitting*.

h. Metrik Evaluasi Rekomendasi Top-N

Evaluasi sistem rekomendasi top-N mengukur kualitas daftar rekomendasi terurut yang disajikan kepada pengguna, bukan akurasi prediksi rating individual. Metrik evaluasi top-N lebih relevan untuk skenario aplikasi praktis di mana sistem menyajikan sejumlah kecil item teratas sebagai rekomendasi (Cremonesi et al., 2010).

Precision@K mengukur proporsi item relevan menggunakan persamaan (2.6):

$$\text{Precision@K} = \frac{| \text{ Relevant} \cap \text{Recommended@K } |}{K} \quad (2.6)$$

di mana Relevant adalah himpunan item relevan bagi pengguna, dan Recommended@K adalah K item teratas dalam daftar rekomendasi. Precision@K mengukur ketepatan rekomendasi namun tidak mempertimbangkan berapa banyak dari semua item relevan yang berhasil ditemukan.

Recall@K mengukur proporsi item relevan yang berhasil ditemukan melalui persamaan (2.7):

$$\text{Recall@K} = \frac{|\ \text{Relevant} \cap \text{Recommended@K}\ |}{|\ \text{Relevant}\ |} \quad (2.7)$$

Recall@K mengukur coverage terhadap seluruh item relevan. Kombinasi Precision dan Recall memberikan gambaran lengkap tentang kualitas rekomendasi (Herlocker et al., 2004).

Normalized Discounted Cumulative Gain (NDCG@K) dihitung menggunakan persamaan (2.8) dan persamaan (2.9):

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)} \quad (2.8)$$

$$\text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}} \quad (2.9)$$

di mana rel_i adalah relevansi item pada posisi i (biasanya 1 untuk relevan dan 0 untuk tidak relevan), dan IDCG@K adalah DCG ideal yang dihitung dari ranking sempurna. NDCG@K dinormalisasi ke rentang [0, 1] sehingga dapat dibandingkan antar pengguna dan *dataset*.

Hit Ratio@K mengukur proporsi pengguna yang memiliki setidaknya satu item menggunakan persamaan (2.10):

$$\text{Hit Ratio@K} = \frac{|\ \{u \in U : |\ \text{Relevant}_u \cap \text{Recommended}_u@K\ | > 0\}\ |}{|\ U\ |} \quad (2.10)$$

di mana U adalah himpunan pengguna dalam *dataset* evaluasi. Hit Ratio@K memberikan gambaran tentang persentase pengguna yang "terlayani" dengan baik oleh sistem rekomendasi.

i. *Dataset Benchmark* dalam Penelitian Sistem Rekomendasi

Dataset benchmark berfungsi sebagai sarana evaluasi yang memungkinkan perbandingan performa antar metode secara objektif. MovieLens adalah salah satu

dataset benchmark yang paling banyak digunakan dalam riset sistem rekomendasi, menyediakan data interaksi pengguna-film dengan berbagai skala mulai dari 100 ribu hingga jutaan rating (Harper & Konstan, 2016)

MovieLens dikembangkan oleh GroupLens Research dan telah menjadi standar *de facto* dalam evaluasi sistem rekomendasi film selama lebih dari dua dekade. *Dataset* ini mencakup rating eksplisit pada skala 1-5 bintang, timestamp interaksi, serta tag yang diberikan pengguna. Ketersediaan data yang konsisten dan tervalidasi membuat MovieLens ideal untuk replikasi eksperimen dan perbandingan metode (Harper & Konstan, 2016).

Dataset MovieLens dapat diperkaya dengan berbagai atribut tambahan dari sumber eksternal seperti The Movie Database (TMDb), yang menyediakan metadata visual (poster), tekstual (sinopsis, overview), dan terstruktur (genre, cast, popularity score). Integrasi MovieLens dengan TMDb memungkinkan penelitian multimodal yang memanfaatkan berbagai sumber informasi untuk membentuk representasi item yang lebih kaya (The Movie Database (TMDB), n.d.). Penggunaan *dataset benchmark* yang telah diperkaya mendukung validasi empiris terhadap pendekatan multimodal dalam kondisi yang lebih realistis.

2.1.1. Pustaka

Sistem rekomendasi memodelkan preferensi pengguna untuk menyaring informasi dalam jumlah besar melalui tiga paradigma utama: *content-based filtering*, *collaborative filtering*, dan *hybrid*. *Content-based filtering* menghasilkan rekomendasi dengan menilai kemiripan fitur item terhadap riwayat ketertarikan pengguna. *Collaborative filtering* memanfaatkan pola kolektif antar pengguna atau antar item tanpa memerlukan informasi konten secara eksplisit (Sarwar et al., 2001). Pendekatan *hybrid* menggabungkan kedua paradigma tersebut untuk mengurangi keterbatasan terkait masalah *cold-start* dan sparsitas pada matriks interaksi, yang pada praktiknya dapat mencapai tingkat sangat tinggi pada *dataset real-world* (Rodríguez-Baena et al., 2025; H. Zhou et al., 2023).

Multimodal learning dalam konteks sistem rekomendasi menggabungkan berbagai sumber informasi untuk memperkaya representasi item dan meningkatkan kualitas prediksi preferensi (Deldjoo et al., 2021). Berbagai sinyal konten seperti

visual, tekstual, dan atribut struktural dapat saling melengkapi dalam merepresentasikan karakteristik item secara lebih lengkap (Liu et al., 2025). Penggunaan model representasi berbasis Transformer untuk ekstraksi fitur multimodal serta integrasinya dengan pendekatan graf kolaboratif seperti LightGCN telah banyak dikaji dalam literatur sebagai strategi untuk membangun sistem rekomendasi yang lebih *robust* terhadap kompleksitas data dan keterbatasan (X. He et al., 2020; Kim et al., 2023; Xia et al., 2024). Perkembangan terkini menunjukkan bahwa arsitektur hibrida yang mengintegrasikan representasi multimodal dengan pemodelan graf secara *end-to-end* mampu meningkatkan performa rekomendasi hingga 10-15% pada skenario *cold-start* dibandingkan pendekatan yang memproses kedua komponen secara terpisah (Klimashevskaia et al., 2024; Liang et al., 2024). Namun, tantangan utama yang masih dihadapi adalah kompleksitas komputasi dan kesulitan dalam optimasi parameter ketika jumlah modalitas bertambah (Lee et al., 2024).

Evaluasi sistem rekomendasi memerlukan *dataset benchmark* yang terstandarisasi untuk memungkinkan perbandingan performa antar metode secara objektif. MovieLens merupakan salah satu *dataset benchmark* yang paling banyak digunakan dalam riset sistem rekomendasi, menyediakan data interaksi pengguna-film yang dapat diperkaya dengan atribut tambahan dari sumber eksternal seperti TMDb (Harper & Konstan, 2016). Kombinasi data interaksi historis dengan metadata multimodal terbukti efektif dalam mengurangi dampak masalah *cold-start* dan meningkatkan akurasi rekomendasi *top-N* pada berbagai skenario evaluasi (Klimashevskaia et al., 2024; Mu & Wu, 2023).

Meskipun berbagai pendekatan multimodal dan graf kolaboratif telah dikembangkan, terdapat *gap* penelitian dalam integrasi *end-to-end* antara representasi konten beragam dengan pemodelan struktur kolaboratif. *Gap* ini berdampak pada ketidakmampuan sistem untuk memanfaatkan secara optimal sinergi antara informasi konten dan sinyal kolaboratif, yang mengakibatkan degradasi performa hingga 30-40% pada item *long-tail* dan *cold-start* yang memiliki interaksi terbatas (Rodríguez-Baena et al., 2025). Selain itu, pendekatan yang memisahkan ekstraksi fitur multimodal dengan pembelajaran graf kolaboratif cenderung mengalami *inconsistency* dalam representasi karena optimasi parameter

dilakukan secara independen, bukan dalam satu kerangka terintegrasi (Liang et al., 2024; Xia et al., 2024). Mengatasi *gap* ini penting untuk membangun sistem rekomendasi yang tidak hanya akurat untuk item populer, namun juga mampu memberikan rekomendasi berkualitas tinggi untuk seluruh spektrum popularitas item, termasuk konten *long-tail* yang seringkali lebih relevan untuk kebutuhan personalisasi mendalam (Lee et al., 2024).

2.1.2. Obyek Penelitian

Obyek dalam penelitian ini adalah sistem rekomendasi film yang dievaluasi menggunakan *dataset* publik MovieLens varian 100K dan 1M yang telah diperkaya dengan metadata dari The Movie Database (TMDb). MovieLens varian 100K mencakup sekitar 100 ribu rating dari 943 pengguna terhadap 1.682 film, sedangkan varian 1M mencakup sekitar 1 juta rating dari 6.040 pengguna terhadap 3.706 film (*GroupLens*, n.d.; Harper & Konstan, 2016). *Dataset* menyediakan data interaksi eksplisit berupa rating skala 1-5 bintang, dilengkapi informasi *timestamp* interaksi serta file pemetaan yang memungkinkan integrasi dengan sumber metadata eksternal. Pengayaan metadata dari TMDb mencakup poster film, sinopsis, informasi *cast*, dan indikator popularitas, yang mendukung pembentukan lingkungan evaluasi multimodal yang relevan untuk penelitian sistem rekomendasi (The Movie Database (TMDB), n.d.).

Dalam konteks penelitian ini, pengguna dimodelkan sebagai entitas yang memberikan rating terhadap film, sedangkan film direpresentasikan sebagai item dengan berbagai modalitas informasi. Representasi item mencakup konten visual dari poster, konten tekstual dari sinopsis, serta metadata terstruktur seperti informasi *cast* dan skor popularitas. Meskipun data rating bersifat eksplisit, penelitian ini mengadopsi paradigma *implicit feedback* (data interaksi tidak langsung seperti pemberian rating) dengan memperlakukan rating sebagai interaksi positif untuk mendukung evaluasi rekomendasi top-N (Rendle et al., 2009).

Tugas utama sistem adalah memprediksi preferensi pengguna terhadap film yang belum pernah berinteraksi serta menyusun daftar rekomendasi top-N yang paling relevan. Evaluasi dilakukan pada berbagai skenario termasuk kondisi *sparsity* yang sangat tinggi dan item *cold-start* dengan interaksi terbatas (Mu & Wu,

2023; Xia et al., 2024). Selain itu, sistem juga dievaluasi melalui stratifikasi berdasarkan lima kategori popularitas film untuk mengukur kemampuan model dalam menangani item populer maupun *long-tail* (Klimashevskaia et al., 2024).

Pemilihan kombinasi MovieLens dan TMDb didasarkan pada perannya sebagai *benchmark* yang banyak digunakan dalam literatur sistem rekomendasi (Harper & Konstan, 2016). Ketersediaan data yang relatif stabil serta kemudahan replikasi memungkinkan perbandingan performa secara objektif dengan pendekatan *state-of-the-art*. Penggunaan *dataset benchmark* ini mendukung validasi empiris terhadap efektivitas arsitektur hibrida yang diusulkan dalam menangani masalah rekomendasi film berbasis konten multimodal dan sinyal kolaboratif.

2.1.3. Metode, Teori, dan Algoritma

Pendekatan multimodal berbasis *Convolutional Neural Network* (CNN) yang digunakan dalam beberapa penelitian sebelumnya memiliki keterbatasan dalam menangkap hubungan global pada representasi visual karena sifat *receptive field* yang bersifat lokal. Mekanisme penggabungan fitur multimodal pada pendekatan tersebut umumnya masih menggunakan konkatenasi sederhana atau *late fusion* melalui *layer fully connected*, sehingga interaksi antar sumber informasi belum dimodelkan secara optimal (Mu & Wu, 2023). Pada domain *collaborative filtering* berbasis graf, pendekatan seperti *Neural Graph Collaborative filtering* menambahkan transformasi linear dan fungsi aktivasi nonlinear pada setiap *layer* propagasi graf, yang dalam praktiknya tidak selalu memberikan peningkatan performa signifikan dan justru meningkatkan kompleksitas komputasi (Wang et al., 2019).

LightGCN mengatasi keterbatasan tersebut dengan menyederhanakan proses pembelajaran graf melalui penghapusan seluruh transformasi linear dan fungsi aktivasi nonlinear, serta hanya mempertahankan operasi agregasi *neighbor* secara linear pada graf bipartit pengguna-item (He et al., 2020). Propagasi *embedding* dilakukan dengan cara mengagregasi *embedding* tetangga pada graf, di mana *embedding* pengguna pada *layer* berikutnya dihitung dari rata-rata tertimbang *embedding* item yang berinteraksi dengannya, dan sebaliknya. Normalisasi simetris

digunakan untuk memastikan stabilitas numerik selama propagasi sinyal, sedangkan *embedding* akhir diperoleh melalui agregasi dari semua *layer* propagasi dengan bobot yang dapat dipelajari. Penyederhanaan ini secara signifikan mengurangi jumlah parameter dan beban komputasi, sekaligus terbukti efektif dalam menyebarkan sinyal preferensi antar node, sehingga LightGCN menjadi *backbone* yang banyak digunakan dalam sistem *collaborative filtering* modern karena efisiensi dan stabilitasnya (Lee et al., 2024). Studi terkini menunjukkan bahwa integrasi LightGCN dengan mekanisme *knowledge-aware attention* dapat meningkatkan interpretabilitas model dan performa pada skenario *cold-start* melalui penggabungan sinyal struktural dan semantik (Pujahari et al., 2026).

Model *Transformer* digunakan sebagai *feature extractor* karena kemampuannya menangkap dependensi jarak jauh melalui mekanisme *self-attention*. *Vision Transformer* efektif dalam memodelkan hubungan global pada representasi visual dengan membagi gambar poster menjadi *patches* dan memodelkan interaksi antar *patches*, sementara BERT menghasilkan representasi tekstual yang kaya semantik dan kontekstual dari sinopsis film (Palanisamy et al., 2025). Integrasi representasi dari berbagai modalitas dilakukan melalui mekanisme *cross-attention* yang mempelajari korespondensi antar sumber informasi secara adaptif, di mana modalitas visual dapat "bertanya" kepada modalitas tekstual tentang informasi relevan dan sebaliknya (Xia et al., 2024). Mekanisme ini menghitung skor perhatian berdasarkan kesamaan antara *query* dari satu modalitas dan *key* dari modalitas lain, kemudian membentuk representasi output melalui agregasi tertimbang dari *value* berdasarkan skor perhatian tersebut. Pendekatan ini memungkinkan pembentukan *embedding* item yang lebih informatif dibandingkan konkatenasi sederhana atau *late fusion* (Álvarez-Sánchez et al., 2025). Penelitian mutakhir menunjukkan bahwa *cross-attention* berbasis Transformer mampu memfasilitasi transfer informasi inter-modal dan intra-modal secara simultan, menghasilkan representasi yang lebih koheren untuk sistem rekomendasi multimodal (Li et al., 2024).

Berdasarkan pertimbangan tersebut, penelitian ini mengembangkan arsitektur hibrida yang mengintegrasikan representasi konten multimodal berbasis Transformer dengan *embedding* kolaboratif item pada LightGCN melalui

mekanisme aditif. *Embedding* multimodal yang dihasilkan dari fusi fitur visual, tekstual, dan metadata terstruktur dikombinasikan dengan *embedding* kolaboratif item menggunakan parameter α yang dipelajari secara adaptif, sehingga bobot kontribusi informasi multimodal dapat disesuaikan oleh model selama proses *training*. *Embedding* pengguna diinisialisasi secara *random* sebagai parameter yang dipelajari dari pola interaksi historis. Dengan mekanisme ini, informasi konten dari berbagai modalitas memperkaya proses propagasi sinyal kolaboratif pada graf bipartit tanpa menghilangkan sinyal interaksi yang sudah ada.

Proses pembelajaran dilakukan secara *end-to-end* dengan fungsi objektif *Bayesian Personalized Ranking* (BPR) yang dirancang untuk pembelajaran dari *implicit feedback*. BPR mengoptimalkan model dengan paradigma *pairwise ranking*, di mana sistem belajar untuk memberikan skor lebih tinggi pada item yang berinteraksi dengan pengguna dibandingkan item yang tidak berinteraksi, dengan regularisasi untuk mencegah *overfitting* (Rendle et al., 2009). Seluruh parameter model termasuk parameter Transformer untuk ekstraksi fitur multimodal dan parameter LightGCN untuk propagasi graf dilatih secara bersama-sama, memungkinkan informasi konten dan sinyal kolaboratif saling memperkaya dalam satu kerangka terintegrasi.

Pendekatan ini ditujukan untuk meningkatkan kualitas representasi item, khususnya pada skenario *cold-start* di mana item memiliki sedikit atau bahkan tidak ada data interaksi, serta item *long-tail* yang jarang berinteraksi dengan pengguna. Integrasi *end-to-end* antara representasi konten multimodal dan pemodelan graf kolaboratif mengatasi tantangan yang masih menjadi limitasi pada pendekatan multimodal dan berbasis graf apabila digunakan secara terpisah (Lee et al., 2024; Liang et al., 2024; Xia et al., 2024).

2.2 Penelitian Terdahulu

Bagian ini menyajikan ringkasan dan analisis beberapa penelitian sebelumnya yang berkaitan dengan sistem rekomendasi film, pendekatan rekomendasi multimodal, serta metode rekomendasi berbasis graf. Studi yang dirangkum mencakup penelitian yang fokus pada pemodelan konten multimodal tanpa graf, pemodelan graf kolaboratif tanpa pemanfaatan konten secara eksplisit,

serta pendekatan yang mulai menggabungkan keduanya. Literatur yang disertakan berasal dari publikasi ilmiah terindeks dan dapat diakses secara luas, termasuk beberapa karya yang relevan sebagai baseline eksperimen dalam penelitian ini. Selain itu, ditinjau pula studi terkini pada rentang tahun 2023–2025 yang mengkaji multimodal deep learning dan graph-based recommendation, sehingga posisi dan kontribusi penelitian yang diusulkan dapat dipetakan secara jelas dibandingkan pendekatan sebelumnya.

Tabel 2. 1 Penelitian Terdahulu

Peneliti/Judul/Tahun	Metode	*Dataset*	Keterbatasan	Relevansi
(Mu & Wu, 2023) Multimodal Movie Recommendation System Using Deep Learning	CNN untuk ekstraksi fitur poster, dense *layer* untuk metadata terstruktur (genre), dan matrix factorization.	*Dataset* MovieLens 100K dan 1M	Menggunakan CNN dan matrix factorization klasik; belum mengeksplorasi graph neural network dan arsitektur Transformer modern	Menjadi referensi langsung untuk sistem rekomendasi film multimodal berbasis poster di *dataset* yang sama
(Kim et al., 2023) MONET: Modality-Embracing Graph Convolutional	Graph Convolutional Network multimodal dengan mekanisme target-	*Dataset* multimedia recommendation (Amazon)	Kompleksitas arsitektur tinggi; tidak menggunakan ViT dan BERT sebagai ekstraktor fitur; fokus	Menjadi referensi pendekatan graph multimodal dengan attention; mendukung

Peneliti/Judul/Tahun	Metode	*Dataset*	Keterbatasan	Relevansi
Network and Target-Aware Attention for Multimedia Recommendation	aware attention untuk integrasi modalitas visual, teks, dan audio		pada multimedia umum, bukan domain film	justifikasi integrasi GNN dan mekanisme attention dalam penelitian ini
(Liu et al., 2025) Multimodal Recommender Systems: A Survey	Survei komprehensif mencakup taksonomi arsitektur, modalitas, skema pelatihan, dan tantangan riset sistem rekomendasi multimodal	*Dataset* rekomendasi dengan rating dan side information	Bersifat survei, tidak menawarkan model baru yang dapat dijadikan baseline eksperimen langsung	Menjadi landasan teoritis untuk memposisikan penelitian ini dalam lanskap riset multimodal terkini dan mengidentifikasi gap penelitian
(Lee et al., 2024) *Embedding* Enhancement Method for	Teknik peningkatan *embedding* pada LightGCN	*Dataset* rekomendasi dengan rating dan side information	Bergantung pada struktur LightGCN murni; tidak mengintegrasikan	Menjadi referensi untuk optimasi komponen LightGCN;

Peneliti/Judul/Tahun	Metode	*Dataset*	Keterbatasan	Relevansi
LightGCN in Recommendation Information Systems	untuk mengurangi degradasi representasi pada lapisan yang dalam		representasi multimodal berbasis konten secara eksplisit	mendukung pemilihan LightGCN sebagai backbone *collaborative filtering*
(Xia et al., 2024) Movie Recommendation with Poster Attention via Multi-modal Transformer Feature Fusion	Multimodal Transformer dengan modul Poster Attention untuk fusi fitur visual poster dan sinopsis	*Dataset* MovieLens 100K dan 1M	Belum mengintegrasikan LightGCN sebagai komponen *collaborative filtering* berbasis graf; evaluasi belum mencakup skenario *cold-start* secara eksplisit	Menjadi baseline langsung penelitian ini karena menggunakan *dataset* dan domain yang sama; gap-nya diatasi dengan penambahan LightGCN
(Álvarez-Sánchez et al., 2025) Leveraging Visual Side Information in	Vision Transformer (ViT) sebagai ekstraktor fitur visual berkualitas	Beberapa *dataset* rekomendasi dengan fitur gambar produk	Fokus pada modalitas visual saja; tidak menggabungkan modalitas teks (sinopsis)	Membuktikan keunggulan ViT dibandingkan CNN untuk ekstraksi fitur visual dalam

Peneliti/Judul/Tahun	Metode	*Dataset*	Keterbatasan	Relevansi
Recommender Systems via Vision Transformer Architectures	tinggi sebagai side information dalam sistem rekomendasi		secara eksplisit dalam representasi item	rekomendasi; menjadi justifikasi utama pemilihan ViT pada penelitian ini
(Liang et al., 2024) Lightweight Embeddings for Graph *Collaborative filtering*	Teknik *embedding* ringan untuk graph-based *collaborative filtering* yang mengurangi kompleksitas komputasi	Berbagai *dataset benchmark collaborative filtering*	Belum mengintegrasikan informasi multimodal secara eksplisit; potensi kombinasi dengan fitur visual dan teks masih belum dieksplorasi	Menjadi referensi untuk efisiensi komputasi komponen LightGCN; mendukung kelayakan implementasi arsitektur hibrida pada *dataset* skala menengah
(Li et al., 2024) A Multimodal Graph Recommendation Method Based	Metode rekomendasi graf multimodal dengan mekanisme *cross-*	*Dataset* rekomendasi multimodal (Amazon)	Tidak menggunakan ViT dan BERT sebagai ekstraktor fitur; belum dievaluasi	Menjadi referensi utama untuk justifikasi pemilihan *cross-attention*

Peneliti/Judul/Tahun	Metode	*Dataset*	Keterbatasan	Relevansi
on *Cross-attention* Fusion	*attention* untuk fusi representasi berbagai modalitas		pada domain film dengan *dataset* MovieLens	sebagai modul fusi; menunjukkan efektivitas kombinasi graph recommendation dan *cross-attention*
(Xu et al., 2025) COHESION: Composite Graph Convolutional Network with Dual-Stage Fusion for Multimodal Recommendation	Composite Graph Convolutional Network dengan dual-stage fusion untuk rekomendasi multimodal	*Dataset* rekomendasi multimodal benchmark	Arsitektur dual-stage fusion lebih kompleks dan belum dievaluasi secara spesifik pada domain film dengan *dataset* MovieLens	Menunjukkan bahwa integrasi graph convolutional network dan fusi multimodal merupakan tren riset mutakhir; memvalidasi arah penelitian ini sejalan dengan perkembangan terkini SIGIR 2025

Peneliti/Judul/Tahun	Metode	*Dataset*	Keterbatasan	Relevansi
(LeeGeon et al., 2025) Revisiting LightGCN: Unexpected Inflexibility, Inconsistency, and A Remedy Towards Improved Recommendation	Analisis kritis dan perbaikan arsitektur LightGCN untuk meningkatkan fleksibilitas dan konsistensi model	Berbagai *dataset* rekomendasi berbasis graf	Fokus pada perbaikan struktural LightGCN murni; tidak mengintegrasikan representasi multimodal berbasis Transformer	Menjadi justifikasi penggunaan LightGCN sebagai komponen yang telah terbukti dan terus dikembangkan komunitas riset; mendukung fondasi teoritis arsitektur hibrida

2.3 Analisis Kebaruan Penelitian

Penelitian sebelumnya menunjukkan dua jalur pengembangan yang umumnya masih berjalan terpisah. Jalur pertama fokus pada representasi konten multimodal menggunakan arsitektur Transformer untuk menggabungkan fitur visual dan tekstual (Mu & Wu, 2023; Xia et al., 2024), namun belum mengintegrasikan komponen graf kolaboratif untuk menangkap pola preferensi kolektif. Jalur kedua mengembangkan pemodelan graf seperti LightGCN yang efisien dalam menyebarkan sinyal kolaboratif (X. He et al., 2020; LeeGeon et al., 2025), namun belum memanfaatkan informasi konten multimodal secara eksplisit. Meskipun MONET (Kim et al., 2023) mulai mengintegrasikan multimodal dengan

GCN, kompleksitas arsitekturnya tinggi dan belum dioptimasi khusus untuk domain film.

Kebaruan utama penelitian ini terletak pada perancangan arsitektur hibrida yang mengintegrasikan representasi multimodal berbasis Transformer dengan mekanisme propagasi graf LightGCN dalam satu kerangka *end-to-end*. Berbeda dengan konkatenasi sederhana atau pemrosesan terpisah, pendekatan ini mengombinasikan *embedding* multimodal dari fusi *cross-attention* secara aditif dengan *embedding* kolaboratif item melalui parameter α yang dipelajari secara adaptif, sehingga sinyal konten dan sinyal kolaboratif saling memperkaya tanpa salah satunya mendominasi. Integrasi ini ditujukan untuk mengatasi limitasi pendekatan yang hanya mengandalkan satu aspek, khususnya dalam menangani skenario *cold-start* dan item *long-tail* yang masih menjadi tantangan utama.

Selain aspek arsitektural, penelitian ini juga berkontribusi pada penyediaan *pipeline* pengolahan *dataset* publik yang dapat direplikasi untuk penelitian sistem rekomendasi film. Evaluasi dilakukan secara komprehensif dengan studi ablasi untuk mengukur kontribusi spesifik setiap komponen (ekstraksi visual, ekstraksi tekstual, fusi multimodal, propagasi graf) terhadap peningkatan performa rekomendasi top-N pada *dataset* MovieLens varian 100K dan 1M yang telah diperkaya dengan metadata TMDb.