BAB 2
LANDASAN TEORI

Pada bagian ini dijabarkan teori-teori, literatur, dan konsep fundamental yang mendasari perancangan dan pengembangan sistem peringkas video (*clipping*) berbasis *Multimodal Deep Learning*. Fokus utama literatur bertumpu pada pemanfaatan arsitektur kecerdasan buatan untuk mendeteksi momen penting (*highlight*) pada tayangan siaran langsung (*live streaming*) YouTube Indonesia, dengan mengekstraksi dan memadukan parameter video dari tiga modalitas utama agar saling melengkapi.

2.1 Tinjauan Teori

Bagian ini memaparkan landasan teoretis mengenai konsep dasar dan komponen-komponen utama yang digunakan dalam pengembangan sistem peringkas video multimodal.

2.1.1 Peringkasan Video dan Deteksi Momen Penting

Peringkasan video (*video summarization*) merupakan proses penyulingan tayangan berdurasi panjang menjadi rangkuman pendek (klip) yang merepresentasikan bagian paling krusial, emotif, atau penting dari keseluruhan durasi [3]. Pendekatan tradisional umumnya melakukan ekstraksi dengan mengandalkan perubahan tingkat kecerahan layar, perpindahan transisi kamera (*scene cut*), maupun pengenalan objek khusus secara modalitas tunggal [3].

Namun, seiring dengan berkembangnya variasi tipe tayangan, khususnya penyiaran langsung (*live stream*), metode tersebut dinilai tidak efektif [2]. Konten *streaming* umumnya mempertahankan sudut kamera statis, namun memiliki volatilitas informasi yang sangat tinggi pada respons audiens dan suara pembicara. Oleh karena itu, deteksi momen penting tidak lagi dapat diandalkan hanya pada satu aspek [2]. Ekstraksi *highlight* mutakhir kini diukur tidak sekadar dari kualitas gambar, melainkan kemampuan sistem untuk meniru heuristik (*heuristic*) manusia dalam menilai antusiasme temporal sebuah siaran secara utuh.

2.1.2 Representasi Vektor dan Ruang Vektor (Vector Space Embedding)

Dalam arsitektur pembelajaran mesin (*machine learning*) dan *deep learning* multimodal, data mentah dari berbagai modalitas—seperti teks percakapan penonton, bingkai gambar video, dan sinyal gelombang suara audio—tidak dapat diproses secara langsung oleh jaringan saraf tiruan (*neural network*) dalam format mentahnya. Model kecerdasan buatan secara fundamental hanya dapat melakukan perhitungan pada data numerik (angka). Oleh karena itu, seluruh masukan multimedia dari berbagai jenis sumber data harus ditransformasikan terlebih dahulu menjadi representasi vektor numerik berdimensi tinggi (*vector embedding*).

Vektor $\mathbf{v} \in \mathbb{R}^D$ merupakan susunan *array* bilangan riil berukuran konstan D. Pada penelitian ini, seluruh modalitas diekstraksi menjadi vektor berdimensi konstan $D = 768$, yang meliputi:

1. Vektor Teks: Konteks pesan obrolan langsung (*live chat*) diproses oleh model IndoBERT untuk menghasilkan vektor numerik berukuran 768 yang mengekstraksi makna semantik dan sentimen penonton.

2. Vektor Visual: Bingkai gambar (*frame*) video per detik diproses oleh model CLIP untuk menghasilkan vektor numerik berukuran 768 yang mengekstraksi fitur spasial dan Objek visual.

3. Vektor Audio: Sinyal gelombang suara per detik diproses oleh model AST untuk menghasilkan vektor numerik berukuran 768 yang mengekstraksi intensitas akustik dan dinamika nada.

Transformasi ke dalam bentuk vektor numerik ini menyajikan tiga fungsi mendasar dalam arsitektur sistem:

1. Penyeragaman Dimensi Data: Mengubah data masukan yang berasal dari jenis sumber berbeda (teks, gambar, dan audio) menjadi bentuk representasi numerik dengan ukuran dimensi yang seragam ($D = 768$).

2. Ruang Vektor dan Kemiripan Semantik (*Vector Space*): Di dalam ruang laten (*latent space*) 768 dimensi ini, sampel-sampel atau konsep yang memiliki kemiripan emosi atau konteks akan dipetakan pada titik koordinat yang saling berdekatan secara matematis (diukur menggunakan kemiripan kosinus / *cosine similarity* atau perkalian titik / *dot product*).

3. Memungkinkan Operasi Fusi Multimodal: Karena seluruh modalitas telah berdimensi seragam ($D = 768$), model jaringan saraf (TripleSumm) dapat menjalankan operasi aljabar linier seperti penggabungan (*concatenation*) dan pembobotan atensi silang (*cross-attention fusion*) secara serentak untuk mendeteksi momen penting.

2.1.3 Pelabelan Otomatis (Weakly-Supervised Labeling)

Salah satu tantangan utama dalam membangun sistem deteksi momen penting adalah penyediaan label acuan otomatis (*pseudo-label*). Anotasi manual oleh manusia pada video berdurasi panjang sangat mahal dan sulit diterapkan pada jumlah data yang besar. Penelitian ini mengadopsi pendekatan pelabelan otomatis (*weakly-supervised*) dengan menggabungkan dua sinyal proksi (*proxy signal*) [4, 2].

A Local Rolling Z-Score Volume Obrolan

Metode ini memanfaatkan volume pesan obrolan (*message count*) dalam setiap jendela waktu (*sliding window*) sebagai indikator tingkat antusiasme penonton [2]. Terinspirasi dari pendekatan penentuan puncak (*peak detection*) obrolan pada sistem CatchLive [4], penelitian ini mengembangkan rumusan secara mandiri untuk menghitung ambang batas lokal. Alih-alih menggunakan statistik global yang rentan terhadap bias segmen pembuka (*intro*) dan penutup (*outro*), penelitian ini mendefinisikan perhitungan ambang batas lokal berdasarkan aliran temporal menggunakan *rolling mean* dan *rolling standard deviation* [10]:

$$\text{threshold}_t = \bar{x}_t + \alpha \cdot \sigma_t \tag{2.1}$$

di mana $\bar{x}_t$ dan σ_t masing-masing adalah rata-rata dan simpangan baku volume obrolan pada jendela berpusat di waktu t, serta α adalah faktor pengali sensitivitas [2, 4].

B Intensitas YouTube Heatmap (Most Replayed)

YouTube Heatmap merupakan data statistik resmi dari platform YouTube yang merekam segmen-segmen video yang paling sering diulang-putar oleh penonton secara agregat [5]. Data ini diekstrak menggunakan *library* `yt-dlp`

dan diinterpolasi secara temporal ke jendela waktu yang sesuai. Segmen dengan intensitas *heatmap* di bawah ambang batas tertentu ($\geq$ 0,2) disaring untuk mereduksi alarm palsu (*false positive*) [5].

Pelabelan akhir menggabungkan kedua sinyal tersebut: suatu segmen dilabeli sebagai momen penting (*highlight*) jika dan hanya jika volume obrolan melebihi ambang batas lokal dan intensitas *heatmap* memenuhi ambang batas minimum [5].

2.1.4 Pemrosesan Bahasa Alami dan IndoBERT

Pemrosesan Bahasa Alami (*Natural Language Processing* – NLP) adalah sub-bidang dari kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Pada ekosistem tayangan *streaming*, sumber utama teks didapat dari rekam jejak obrolan penonton (*live chat*) yang melintas secara *real-time* [11].

A Arsitektur BERT

Bidirectional Encoder Representations from Transformers (BERT) merupakan model representasi bahasa pra-latih (*pre-trained*) yang diperkenalkan oleh Devlin dkk. [12]. Berbeda dengan model bahasa konvensional yang membaca teks secara searah (kiri-ke-kanan atau kanan-ke-kiri), BERT menggunakan pendekatan dua arah (*bidirectional*) sehingga mampu memahami konteks kata dari kedua sisi kalimat secara serentak. Arsitektur BERT dibangun dari tumpukan lapisan *Transformer Encoder* dan dilatih menggunakan dua tugas (*pre-training task*):

- *Masked Language Model (MLM):* Sebagian token dalam kalimat masukan disembunyikan secara acak, dan model dilatih untuk memprediksi token yang disembunyikan tersebut berdasarkan konteks sekitarnya.
- *Next Sentence Prediction (NSP):* Model dilatih untuk memprediksi apakah dua kalimat yang diberikan merupakan pasangan yang berurutan secara alami atau tidak.

B IndoBERT

IndoBERT merupakan varian BERT yang dilatih secara khusus pada korpus teks bahasa Indonesia berskala besar sebagai bagian dari *benchmark* IndoNLU [6]. Penelitian ini menggunakan varian IndoBERT-base-p2 (`indobenchmark/indobert-base-p2`) yang memiliki spesifikasi arsitektur sebagaimana ditunjukkan pada Tabel 2.1.

Tabel 2.1. Spesifikasi Arsitektur IndoBERT-base-p2

Parameter	**Nilai**
Jumlah Lapisan (*Layers*)	12
Dimensi Tersembunyi (*Hidden Size*)	768
Jumlah *Attention Heads*	12
Total Parameter	$\sim$110 juta
Kosakata (*Vocabulary*)	30.522 token
Metode Tokenisasi	*WordPiece*
Panjang Masukan Maksimum	512 token

Masalah yang sering timbul pada implementasi fusi adalah penggunaan arsitektur bahasa Inggris murni (seperti RoBERTa [13]) yang gagal mengenali semantik dari kata-kata gaul, *slang*, maupun ekspresi sosiokultural audiens Indonesia. Hal ini disebabkan oleh keterbatasan kamus kosakata (*vocabulary*) dari model bahasa Inggris yang tidak memuat entri kosakata informal lokal, sehingga memicu masalah *Out-of-Vocabulary* (OOV). Ketika memproses kata-kata non-baku Indonesia, *tokenizer* berbasis bahasa Inggris akan memecah istilah tersebut secara berlebihan menjadi pecahan karakter yang tidak representatif secara semantik (misalnya kata gaul seperti "gokil" atau "bjir" dipecah menjadi unit-unit sub-kata Inggris yang acak), yang mengakibatkan hilangnya informasi makna kontekstual secara signifikan. Bukti empiris mengenai kesenjangan ini juga ditunjukkan pada pengujian *benchmark* IndoNLU oleh Wilie dkk. [6], di mana model kebahasaan lokal terbukti mengungguli model multibahasa maupun model non-lokal secara signifikan dalam memahami sentimen teks informal berbahasa Indonesia. Oleh karena itu, dilakukan adaptasi pada komponen *encoder* teks dalam arsitektur fusi multimodal dengan mengganti *backbone* bahasa bawaan menggunakan IndoBERT [6]. Melalui tahap *fine-tuning* tambahan menggunakan basis data *chat* dari platform penyedia siaran langsung, model ini diadaptasi ulang agar sensitif terhadap

kepadatan reaksi berlebihan (*hype*) yang identik dengan kehadiran momen klimaks pada video.

2.1.5 Ekstraksi Fitur Audio: Audio Spectrogram Transformer (AST)

Audio Spectrogram Transformer (AST) yang diperkenalkan oleh Gong dkk. [8] merupakan model pendeteksi audio mutakhir yang mengadaptasi arsitektur *Vision Transformer* (ViT) untuk memproses representasi visual dari sinyal audio. Model ini mengonversi pita gelombang suara mentah menjadi citra spektrogram frekuensi Mel (*Mel-Spectrogram*), yang kemudian dipecah menjadi sejumlah *patch* dan diproses oleh lapisan *Transformer Encoder* layaknya sebuah gambar dua dimensi.

Penelitian ini menggunakan varian AST yang telah dipra-latih pada *dataset* AudioSet (`MIT/ast-finetuned-audioset-10-10-0.4593`). Spesifikasi model ditunjukkan pada Tabel 2.2.

Tabel 2.2. Spesifikasi *Audio Spectrogram Transformer* (AST)

Parameter	**Nilai**
Arsitektur *Backbone*	*Vision Transformer* (ViT)
Pre-training Dataset	AudioSet (2,1 juta klip audio)
Laju Sampel Masukan	16.000 Hz
Panjang Segmen Per Jendela	10 detik (*centered*)
Dimensi Keluaran (*Embedding*)	768
Strategi Agregasi	Rata-rata pada dimensi temporal (*mean pooling*)
Status pada Penelitian	*Frozen* (tanpa *fine-tuning*)

Model AST tidak hanya menangkap kata-kata yang diucapkan, namun lebih mengutamakan intensitas akustik, tawa, teriakan, maupun perubahan nada (*pitch*) mendadak yang sering mengindikasikan klimaks emosional [8]. Sebagaimana CLIP, AST juga tidak di-*fine-tune*; fitur audio diekstraksi terlebih dahulu secara terpisah sebelum pelatihan dimulai dan disimpan sebagai berkas `.npy` demi efisiensi komputasi.

2.1.6 Ekstraksi Fitur Visual: Contrastive Language-Image Pre-Training (CLIP)

Contrastive Language-Image Pre-Training (CLIP) yang dikembangkan oleh Radford dkk. [7] merupakan model fondasi (*foundation model*) berbasis arsitektur *dual encoder* yang melatih penyandi gambar (*Image Encoder*) dan penyandi teks (*Text Encoder*) secara bersamaan menggunakan tujuan pelatihan kontrastif (*contrastive learning*). Model ini dilatih pada lebih dari 400 juta pasangan teks-gambar dari internet, sehingga menghasilkan representasi visual yang sangat kaya secara semantik.

Dalam penelitian ini, hanya komponen *Image Encoder* dari CLIP yang digunakan, yaitu varian ViT-L/14 (`openai/clip-vit-large-patch14`). Spesifikasi model ditunjukkan pada Tabel 2.3.

Tabel 2.3. Spesifikasi CLIP ViT-L/14 Image Encoder

Parameter	**Nilai**
Arsitektur *Backbone*	*Vision Transformer* (ViT-L/14)
Ukuran *Patch*	14×14 piksel
Resolusi Masukan	224×224 piksel
Dimensi Keluaran (*Embedding*)	768
Status pada Penelitian	*Frozen* (tanpa *fine-tuning*)

Dibandingkan menggunakan jaringan saraf konvolusional (CNN) standar, CLIP memiliki keunggulan adaptasi karena dipelajari dari jutaan pasangan teks dan gambar sehingga memiliki pemahaman konteks visual yang lebih kaya [7]. CLIP tidak di-*fine-tune* pada penelitian ini; fitur visual diekstraksi terlebih dahulu sebelum proses pelatihan dimulai untuk setiap detik durasi video dan disimpan dalam format berkas `.npy`. Pendekatan ini dilakukan untuk menghindari pemrosesan gambar berulang secara redundan pada setiap *epoch* pelatihan model fusi, sehingga mempercepat waktu komputasi dan menghindari kelebihan beban memori GPU.

2.1.7 Multimodal Deep Learning dan Arsitektur Fusi

Multimodal Deep Learning merupakan cabang pembelajaran mesin (*machine learning*) yang bertugas menggabungkan, menganalisis, dan menarik kesimpulan berdasarkan lebih dari satu tipe struktur data (seperti gabungan teks,

audio, dan gambar) secara simultan. Dalam konteks peringkasan video, pendekatan ini krusial untuk mengatasi kekurangan atau *information loss* ketika salah satu sumber data mengalami anomali (misalnya video tanpa suara, atau video sepi *chat*) [14].

Proses penggabungan fitur dari berbagai modalitas ini disebut dengan fusi (*fusion*). Secara umum, terdapat tiga jenis skema fusi yang umum digunakan [3]:

1. Fusi Awal (*Early Fusion):* Menggabungkan data mentah atau fitur tingkat rendah dari berbagai modalitas sebelum dimasukkan ke dalam jaringan klasifikasi. Kelemahannya adalah kesulitan menyelaraskan data dengan representasi dan dimensi yang sangat berbeda (misalnya, matriks piksel dengan vektor teks).

2. Fusi Akhir (*Late Fusion):* Melatih model independen untuk tiap modalitas dan menggabungkan hasil klasifikasi akhirnya. Meskipun stabil, skema ini gagal menangkap interaksi korelasi antarmodalitas sejak awal.

3. Fusi Menengah / Adaptif (*Intermediate / Adaptive Fusion):* Menggunakan arsitektur jaringan tingkat lanjut yang mengintegrasikan fitur dari berbagai modalitas di tingkat representasi menengah. Model seperti Mosu dari TripleSumm menerima lapisan fitur (*feature maps*) dan melakukan peleburan melalui mekanisme atensi silang (*cross-attention*) [5]. Proses atensi dinamis inilah yang menentukan probabilitas suatu jendela waktu (*window*) apakah layak dilabeli sebagai momen penting. Penelitian ini mengadopsi skema fusi menengah.

A Transfer Learning pada Lapisan Fusi Multimodal

Membangun model fusi *multimodal* dari awal (*from scratch*) memerlukan dataset yang sangat besar dan waktu pelatihan yang lama untuk mempelajari relasi spasial, temporal, dan semantik antar-modalitas secara optimal. Oleh karena itu, konsep *transfer learning* sering diterapkan untuk meningkatkan performa model pada tugas dengan data terbatas. *Transfer learning* adalah metode pembelajaran di mana pengetahuan yang diperoleh saat melatih model untuk suatu tugas didelegasikan untuk menyelesaikan tugas lain yang serupa [9].

Dalam penelitian ini, *transfer learning* diterapkan dengan memuat bobot pra-latih (*pre-trained weights*) dari model TripleSumm [5]. Model TripleSumm asli dilatih menggunakan dataset multimodal skala besar MoSu untuk menyelaraskan

fitur visual (dari CLIP ViT-L/14 [7]), audio (dari AST [8]), dan teks (dari RoBERTa [13]) pada ruang laten bersama berdimensi 128 melalui lapisan fusi atensi silang. Melalui pemuatan bobot pra-latih ini, model dapat langsung mengadopsi kemampuan koordinasi dan interaksi multimodal yang telah terbentuk, sehingga hanya perlu menyesuaikan pemetaan (*projection*) dari representasi *backbone* baru (CLIP ViT-L/14 berdimensi 768, AST berdimensi 768, dan IndoBERT berdimensi 768) ke ruang laten bersama tersebut.

2.1.8 Arsitektur Transformer

Arsitektur *Transformer* yang diperkenalkan oleh Vaswani dkk. [15] telah menjadi fondasi utama dalam berbagai tugas pemrosesan bahasa alami (*Natural Language Processing* – NLP), visi komputer (*Computer Vision*), dan analisis audio. Keunggulan utama *Transformer* terletak pada mekanisme *self-attention* yang memungkinkan model memproses seluruh elemen masukan secara paralel, tanpa dibatasi oleh ketergantungan sekuensial sebagaimana pada jaringan saraf rekuren (*Recurrent Neural Network* / RNN).

A Mekanisme Atensi (Attention Mechanism)

Inti dari arsitektur *Transformer* adalah fungsi *Scaled Dot-Product Attention*. Diberikan matriks *Query* (Q), *Key* (K), dan *Value* (V), perhitungan atensi didefinisikan sebagai [15]:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

di mana d_k adalah dimensi vektor *Key*. Faktor pembagi $\sqrt{d_k}$ berfungsi untuk mencegah nilai perkalian titik (*dot product*) menjadi terlalu besar, yang dapat menyebabkan gradien menghilang saat melewati fungsi *softmax*.

B Atensi Multi-Kepala (Multi-Head Attention)

Alih-alih melakukan satu kali komputasi atensi, arsitektur *Transformer* menjalankan h buah *head* atensi secara paralel, masing-masing dengan proyeksi linier yang berbeda. Keluaran dari seluruh *head* kemudian dikonkatenasi dan diproyeksikan kembali [15]:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \tag{2.3}$$

di mana setiap *head* dihitung sebagai:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.4}$$

dengan $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, dan $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ adalah matriks bobot yang dapat dipelajari.

C Atensi Silang (Cross-Attention)

Dalam konteks fusi multimodal, mekanisme *cross-attention* digunakan untuk mengukur relevansi antarmodalitas [15, 5]. Berbeda dengan *self-attention* di mana Q, K, dan V berasal dari sumber yang sama, pada *cross-attention*, *Query* berasal dari satu modalitas sedangkan *Key* dan *Value* berasal dari modalitas lain. Pada penelitian ini, fitur teks dari IndoBERT berperan sebagai *Query*, sementara fitur visual CLIP dan audio AST berperan sebagai *Key* dan *Value*. Formulasi ini memungkinkan model menentukan seberapa relevan informasi visual dan audio terhadap konteks percakapan yang sedang berlangsung [5].

2.1.9 Fungsi Kerugian: Binary Cross-Entropy with Logits with Positive Weight

Konsep entropi (*entropy*) dalam teori informasi yang dikembangkan oleh Claude Shannon mengukur tingkat ketidakpastian atau keacakan dari suatu variabel acak [16]. Dalam pembelajaran mesin, *cross-entropy* (entropi silang) mengukur tingkat perbedaan atau jarak antara dua distribusi probabilitas: distribusi label target yang sebenarnya $y_i \in \{0,1\}$ dan distribusi probabilitas yang diprediksi oleh model $\hat{y}_i = \sigma(x_i)$. Semakin kecil nilai entropi silang yang dihasilkan, semakin dekat pula distribusi probabilitas prediksi model dengan distribusi label target sebenarnya.

Pada tugas klasifikasi biner dengan ketidakseimbangan kelas (*class imbalance*) yang ekstrem—di mana jumlah segmen non-penting jauh mendominasi segmen penting—fungsi kerugian standar seperti *Binary Cross-Entropy* (BCE) cenderung bias terhadap kelas mayoritas [17]. Untuk mengatasi hal ini, penelitian

ini menggunakan *Binary Cross-Entropy with Logits Loss* dengan pembobotan kelas positif (*positive weight*) [5]:

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot w \cdot \log(\sigma(x_i)) + (1-y_i) \cdot \log(1-\sigma(x_i))] \tag{2.5}$$

di mana:

- N adalah jumlah sampel yang dievaluasi.
- $y_i \in \{0,1\}$ menyatakan label acuan otomatis (*pseudo-label*) untuk sampel ke-i.
- x_i menyatakan logit hasil prediksi model sebelum fungsi aktivasi.
- $\sigma(x_i) = \frac{1}{1+e^{-x_i}}$ menyatakan fungsi aktivasi *sigmoid* untuk mengubah logit menjadi probabilitas kontinu $[0,1]$.
- w menyatakan faktor pengali bobot kelas positif (*positive weight* / `pos_weight`), yang dihitung secara dinamis berdasarkan perbandingan rasio antara jumlah sampel kelas negatif terhadap sampel kelas positif pada dataset latih:

$$w = \frac{N_{\text{negatif}}}{N_{\text{positif}}} \tag{2.6}$$

Penggunaan `pos_weight` ini memberikan penalti kesalahan (*loss penalty*) yang lebih berat pada prediksi salah untuk kelas positif (*False Negative*), sehingga model terdorong untuk menyeimbangkan nilai *Recall* pada kelas minoritas (*highlight*) tanpa merusak kestabilan konvergensi gradien [17, 5].

2.1.10 Metrik Evaluasi Kinerja

Kinerja arsitektur fusi *multimodal* harus dapat dievaluasi secara kuantitatif untuk membuktikan keandalannya dibandingkan model dasar (*baseline*) [14]. Evaluasi dilakukan dengan membandingkan deteksi klip dari model terhadap label acuan otomatis (*pseudo-label*) yang dianotasi secara otomatis [2]. Pengukuran ini didasarkan pada metrik *Confusion Matrix* yang terdiri dari empat komponen:

- *True Positive (TP):* Jumlah segmen yang memang merupakan momen penting dan diprediksi benar oleh sistem.
- *True Negative (TN):* Jumlah segmen non-penting yang diprediksi benar sebagai non-penting.
- *False Positive (FP):* Jumlah segmen non-penting yang salah diprediksi sebagai momen penting.
- *False Negative (FN):* Jumlah segmen yang merupakan momen penting nyata namun terlewatkan oleh sistem.

Dari elemen tersebut, diturunkan tiga parameter utama [14]:

A Precision

Precision mengukur proporsi prediksi momen penting yang benar-benar tepat dari seluruh prediksi positif yang dihasilkan model [14]:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.7}$$

B Recall

Recall mengukur proporsi momen penting aktual yang berhasil ditangkap oleh model [14]:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.8}$$

C F1-Score

F1-Score merupakan rata-rata harmonis dari *Precision* dan *Recall*, sehingga memberikan gambaran keseimbangan antara keduanya. Metrik ini merupakan ukuran utama pada deteksi momen penting karena mampu menghindari bias akibat dominansi salah satu parameter [14]:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.9}$$

Nilai F_1 berkisar antara 0 hingga 1, di mana 1 menunjukkan keselarasan sempurna antara model dan label acuan (*pseudo-label*) [14].

2.1.11 Pipeline Inferensi dan Otomatisasi Penyaluran Model

Model kecerdasan buatan yang telah dilatih didelegasikan ke dalam sebuah *pipeline* inferensi otomatis berbasis Python untuk melokalisasi seluruh alur pemrosesan data multimedia [11]. *Pipeline* ini dirancang secara terstruktur untuk mengabstraksi kompleksitas komputasi berat, seperti tokenisasi teks percakapan, ekstraksi fitur visual CLIP, ekstraksi fitur audio AST, dan fusi temporal lintas modalitas, menjadi satu kesatuan skrip eksekusi yang efisien [14].

Melalui otomatisasi alur pemrosesan ini, pengguna cukup memberikan tautan (*link*) video target atau berkas masukan mentah. Program kemudian akan melakukan proses pengunduhan, pemisahan berkas multimedia, inferensi neural secara paralel, dan fusi representasi laten. Terakhir, sistem akan secara otomatis memotong dan menyimpan cuplikan klip (*clipping*) momen-momen penting yang terdeteksi ke dalam direktori lokal, sehingga mempermudah pemakaian (*usability*) dan menjaga integritas hasil ekstraksi secara lokal [11].

2.2 Penelitian Terdahulu

Penelitian mengenai peringkasan video dan deteksi *highlight* menggunakan pendekatan *Multimodal Deep Learning* telah banyak dilakukan. Beberapa studi terdahulu yang relevan digunakan sebagai landasan dan pembanding untuk menunjukkan letak kebaruan (*novelty*) dari penelitian ini. Tabel 2.4 menyajikan ringkasan dan komparasi dari beberapa penelitian referensi utama.

Tabel 2.4. Perbandingan Penelitian Terdahulu

No	Penelitian / Tahun	Fokus & Metode	Kelebihan	Kekurangan / Kesenjangan
1	Kim, dkk. (2026) [5]	TripleSumm: Arsitektur fusi tiga modalitas adaptif menggunakan mekanisme *cross-attention*.	Mampu menyeimbangkan bobot antara visual, audio, dan teks secara dinamis berdasarkan jendela temporal.	Penggunaan modul teks dasar sering mengalami *domain mismatch* jika berhadapan dengan *slang* atau struktur bahasa informal penonton *streaming*.
2	Rajpal, dkk. (2026) [11]	Peringkasan video instruksional berbantuan panduan label.	Sangat presisi untuk video tutorial yang terstruktur.	Tidak cocok untuk siaran langsung yang dinamis dan tak terprediksi (seperti *gaming*).
3	Marevac, dkk. (2025) [14]	Pembandingan komprehensif algoritma klasifikasi *multimodal* untuk *video summarization*.	Memberikan fondasi empiris mengenai performa berbagai seleksi fitur dan pengklasifikasi komputasi klasik maupun modern.	Belum sepenuhnya mengoptimalkan potensi arsitektur atensi silang mutakhir secara *end-to-end*.

Tabel 2.4. Perbandingan Penelitian Terdahulu (Lanjutan)

No	Penelitian / Tahun	Fokus & Metode	Kelebihan	Kekurangan / Kesenjangan
4	Pennec, dkk. (2025) [18]	Peringkasan video durasi panjang via ekstraksi momen kunci (*key moment extraction*).	Mengurangi redundansi temporal pada tayangan panjang secara efektif.	Bergantung kuat pada transisi visual yang berpotensi menghasilkan hilangnya informasi (*information loss*) pada sudut pandang kamera yang statis.
5	Ringer (2022) [2]	Deteksi *highlight live streaming* menggunakan fitur audio, visual, dan teks.	Memanfaatkan fusi tiga modalitas yang terbukti mengatasi kekurangan model modalitas tunggal.	Analisis teks mengandalkan leksikon bahasa Inggris, sehingga sentimen untuk bahasa lokal belum terakomodasi.
6	Yang, dkk. (2022) [4]	Segmentasi dan deteksi *highlight* siaran langsung menggunakan fusi visual, transkrip, dan obrolan interaktif.	Berjalan secara *real-time* untuk menyajikan ringkasan dan *highlight* bagi penonton terlambat dengan detail fleksibel.	Penyatuan modalitas masih menggunakan fusi aturan heuristik terbobot (*heuristic score-level fusion*) yang memerlukan penyesuaian manual.

Tabel 2.4. Perbandingan Penelitian Terdahulu (Lanjutan)

No	Penelitian / Tahun	Fokus & Metode	Kelebihan	Kekurangan / Kesenjangan
7	Radwan, dkk. (2026) [19]	*Benchmark* SONIC-O1 untuk evaluasi model bahasa besar *multimodal* pada pemahaman audio-video.	Menyediakan standar evaluasi yang tangguh di skenario dunia nyata.	Berfokus pada evaluasi pemahaman model, bukan pada pembangunan sebuah sistem peringkas video produksi massal.
8	Li, dkk. (2026) [20]	*ViSIL*: Evaluasi *information loss* pada sistem peringkas visual *multimodal*.	Metodologi pengukuran kerugian informasi (*information loss*) yang sangat komprehensif.	Berfokus secara parsial pada keselarasan visual, mengabaikan aspek lonjakan akustik dan histeria massa di kolom percakapan.

Berdasarkan komparasi pada Tabel 2.4, mayoritas sistem *highlight detection* yang telah mapan memiliki keterbatasan pada analisis kebahasaan ketika dihadapkan pada skenario sosiokultural lokal, atau masih mengandalkan fusi berbasis aturan heuristik kaku seperti pada CatchLive [4]. Penelitian ini menutupi kesenjangan tersebut dengan mencangkokkan model bahasa khusus, yaitu IndoBERT, yang difokuskan untuk memproses percakapan bahasa Indonesia, kata gaul (*slang*), dan sentimen informal khas penonton *live streaming*, sembari tetap mempertahankan kekuatan fusi dari arsitektur terkemuka seperti TripleSumm berbasis *cross-attention* yang adaptif secara otomatis. Selain itu, penelitian ini juga memanfaatkan *backbone* visual dan audio mutakhir berupa CLIP dan AST yang secara empiris lebih adaptif dibandingkan arsitektur generasi sebelumnya.