SD2025040000234

SCENT: *Semantic Chaptering* dengan *Exponential-Decay* dan *Neural Taxonomy*

Pipeline Multi-Modal Efisien untuk *Comprehensive Video Understanding*

SEMIFINAL

BIG DATA CHALLENGE

2025

SCENT: *Semantic Chaptering* dengan *Exponential-Decay* dan *Neural Taxonomy*

ABSTRAK — Pertumbuhan masif konten video pendek di platform media sosial memicu kebutuhan mendesak akan sistem otomatis yang mampu mengekstraksi pengetahuan dari himpunan data video tersebut. Namun, implementasi sistem *video understanding* yang ada saat ini, yang umumnya mengandalkan Vision-Language Models (**`VLMs`**), menghadapi dilema saat melakukan *deployment*: **`VLMs`** berukuran besar (misalnya, `Qwen 2.5-VL-72B`) menuntut infrastruktur GPU *enterprise* yang mahal, sementara **`VLMs`** yang lebih kecil, meskipun dapat dijalankan pada GPU *consumer-grade*, sering kali mengorbankan performa signifikan pada tugas-tugas yang kompleks. Penelitian ini menawarkan pendekatan **modular** untuk mengoptimalkan *video understanding* multi-modal dengan efisiensi sumber daya. Utamanya adalah *pipeline* multi-modal berbasis *chaptering* yang mengintegrasikan ekstraksi *keyframe*, *scene detection*, dan transkripsi audio. *Pipeline* efisien ini menghasilkan empat *output* simultan: penyusunan taksonomi, peringkasan hierarkis, dan analisis konten multi-dimensi (*content scoring* untuk *civility*, *hostility*, dan *commercial insight*), mencapai pemahaman video yang seimbang.

KATA KUNCI: *video understanding*; *multi-modal*; *chaptering*; *keyframe extraction*; *resource-efficient deployment*

I PENDAHULUAN

Isu sentral penelitian ini adalah **keterbatasan sumber daya komputasi** dalam tugas *video understanding* multi-modal, yang menuntut pemrosesan simultan informasi teks, audio, dan visual. Untuk mendukung *demokratisasi* model *video understanding*, penelitian ini mengusulkan sebuah pipeline yang dirancang spesifik untuk efisiensi tanpa mengorbankan kualitas dan kelengkapan informasi yang diekstraksi.

I.1 Motivasi & Latar Belakang

Large Vision-Language Models (**`VLMs`**) (seperti `Qwen 2.5-VL-72B`) menuntut sumber daya komputasi mahal, minimal 4×A100 80GB VRAM, yang tidak praktis untuk sebagian besar aplikasi yang terkait *video understanding* [1]. Di sisi lain, **`VLMs`** kecil (di bawah 10

miliar parameter) memang dapat digunakan pada GPU *consumer-grade*, tetapi kinerjanya akan menurun drastis saat menangani tugas penalaran yang mendalam. *Large* **VLMs** seperti `Qwen 2.5-VL-72B` memerlukan minimum 384GB VRAM dan membutuhkan 8×A100 atau 8×H100 untuk *deployment* standar [1], sementara model 7B atau 14B memerlukan minimal 24GB VRAM. Model `Phi-3` *vision*, meskipun hanya 3,8B parameter, mencapai 69% pada *MMLU benchmark* dan menunjukkan performa yang sebanding dengan model seperti GPT-3.5, namun tetap mengalami keterbatasan pada *factual knowledge* dan dukungan bahasa terbatas. Penelitian ini mencoba untuk mengusulkan sebuah *pipeline vide understanding* alternatif yang lebih efisien dan juga tetap efektif dalam memberikan hasil ringkasan yang berguna.

I.2 Kajian Pustaka

Video Understanding. *Video understanding* adalah cabang dari kecerdasan buatan dan *computer vision* yang mengajarkan mesin menafsirkan, menganalisis, dan memahami konten, konteks, dan aktivitas dalam video. Sub-tugas utamanya meliputi *video chaptering* [2], *keyframe extraction* [3], *video topic modeling*, dan *video summarization*. Secara kolektif, sub-tugas ini memungkinkan model untuk memahami tindakan, signifikansi (*topic modeling*), dan meringkas konten video secara efisien.

Video Chaptering dan Segmentasi. Chapter-Llama mengusulkan *speech-driven chaptering* yang memanfaatkan ASR untuk segmentasi temporal, mencapai *boundary precision* 0,82 pada video berdurasi sekitar satu jam. Pendekatan ini menggunakan *lightweight speech-guided frame selection strategy* dan memproses video satu jam dalam *single forward pass* [2]. Namun, ketergantungan pada narasi audio berpotensi melewatkan informasi visual penting yang tidak tercakup dalam transkrip.

Keyframe Extraction dengan Scene Detection. `LMSKE` mengembangkan *adaptive clustering* menggunakan `CLIP` *embeddings* untuk mengidentifikasi transisi semantik, meningkatkan F1-*score* sebesar 12 – 15% dibandingkan metode *clustering* tradisional [3]. Namun, `LMSKE` beroperasi secara independen tanpa integrasi *audio modality*, berpotensi menghasilkan *keyframes* yang tidak selaras dengan alur naratif video.

Video Topic Modeling dan Content Taxonomy. Metode *topic discovery* tradisional

seperti `BERTopic` [4] (*unsupervised clustering*) memiliki kelemahan: tidak selalu selaras dengan taksonomi eksternal, sensitif terhadap *hyperparameter*, dan inkonsisten pada video pendek. Solusi alternatif adalah ***direct taxonomy mapping*** berbasis *semantic similarity*, dengan teks di-*encode* dan diklasifikasikan ke taksonomi *predefined* melalui *cosine similarity*. Penggunaan model *asymmetric encoding* (`E5` [5]) dapat meningkatkan presisi klasifikasi hierarkis hingga 8–12%. Untuk konteks temporal video, *exponential-decay pooling* terbukti efektif menjaga informasi lintas segmen sambil mempertahankan *local semantics*.

Multi-Modal Summarization dan Strategi LLM. Penelitian terkini mengeksplorasi *multimodal summarization* yang mengintegrasikan video, teks, dan berbagai modalitas, dengan pendekatan *hierarchical transformer* yang memanfaatkan informasi gerakan untuk *high-frequency sampling*. Untuk pemrosesan dokumen panjang, `LangChain` *framework* [6] mengidentifikasi empat paradigma: *Stuffing* (optimal untuk dokumen pendek namun rentan *information dilution*), *MapReduce* (efektif untuk skalabilitas tetapi kehilangan konteks global), *Refine* (mempertahankan akumulasi konteks dengan beban komputasi lebih tinggi), dan *Map-Rerank*.

I.3 Posisi Penelitian

Research Gap dan Rumusan Masalah. Belum ada *framework* yang mengintegrasikan transkripsi audio dan ekstraksi *keyframe* visual secara seimbang dalam suatu *pipeline* modular untuk video pendek. Penelitian ini menjawab: **Bagaimana merancang *chaptering-based pipeline* yang menghasilkan *video understanding* komprehensif (*semantic chaptering*, *taxonomy mapping*, *hierarchical summarization*, *multi-dimensional content scoring*) dengan keterbatasan sumber daya minimal?**

Penelitian ini mengusulkan pendekatan *chaptering-based multi-modal* tiga tahap yang menggabungkan keunggulan metodologi sebelumnya (Tabel 1). Pendekatan ini mengatasi keterbatasan *speech-driven selection* dengan mengadopsi *scene-based keyframe extraction* dengan *histogram filtering*, dikombinasikan dengan *audio transcription* secara *balanced*.

Tujuan Penelitian. Penelitian ini bertujuan untuk: (1) mengembangkan *semantic chaptering pipeline* yang mengintegrasikan *scene-based keyframe extraction* dengan *times-*

Table 1. Posisi penelitian ini dibandingkan pendekatan existing

Aspek	Large VLM	Small VLM	Chapter-Llama	LMSKE	Penelitian Ini
Deployable (local)	✗	✓	✓	✓	✓
Independent Frame Selection	✗	✗	✓ (speech)	✓ (cluster)	✓
Balanced Multi-modal	✓	✓	✗ (audio)	✗ (visual)	✓
Multi-dim. Content Scoring	✗	✗	✗	✗	✓

tamped audio transcription, (2) mengimplementasikan *multi-output framework* yang menghasilkan *taxonomy construction*, *hierarchical summarization*, dan *multi-dimensional content scoring* (meliputi *civility*, *hostility*, dan *commercial insight*) dalam *single pipeline*, dan (3) mengevaluasi *performance-resource trade-off* dari tiga *summarization strategies* (*Stuffing*, *MapReduce*, *Refine*) pada video pendek (< 5 menit).

Signifikansi dan Manfaat. Kontribusi teoretis meliputi: (1) kajian sistematis *chaptering-based pipeline* yang pertama kali mengintegrasikan *scene-based keyframe extraction* dengan *audio transcription* untuk *balanced multi-modal video understanding*, (2) *framework* yang menghasilkan tiga *output* (*topic modeling*, *summarization*, *multi-dimensional content scoring*) dalam satu *pipeline* efisien, dan (3) analisis *performance-resource trade-off* yang komprehensif. Secara praktikal, *framework* ini mendukung platform media sosial dalam *auto-tagging*, *content discovery*, dan *content moderation* dengan infrastruktur terjangkau.

Batasan. Data berasal dari video pendek (< 5 menit) berbahasa Indonesia dan Inggris. Sebanyak 50 sampel video dianotasi manual sebagai *ground truth* untuk evaluasi.

II METODOLOGI

Bab II menguraikan metode *video understanding* yang efisien. Pembahasannya mencakup dataset (Subbab II.1), ekstraksi fitur (Subbab II.2), teknik *chaptering* dan peringkasan (Subbab II.3), proses mendapatkan *insight* dan analisis (Subbab II.4), dan terakhir metode evaluasi (Subbab II.5).

II.1 Dataset

Penelitian ini menggunakan 1000 video pendek (durasi < 5 menit) dari dua sumber: Google Drive (330 video) dan Instagram (670 video). Sebanyak 50 video dianotasi manual sebagai *ground truth* untuk evaluasi.

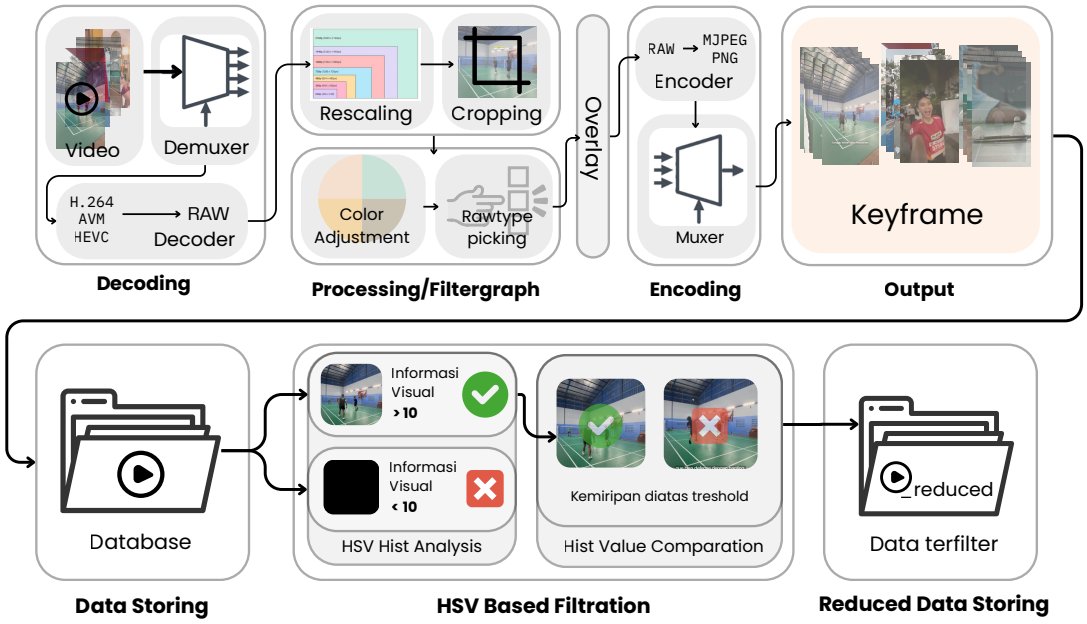


Figure 1. Arsitektur *pipeline* yang diusulkan untuk ekstraksi pengetahuan dan ringkasan video.

II.2 Preprocessing dan Feature Extraction

Visual Processing: Keyframe Extraction. Ekstraksi *keyframe* mengadopsi pendekatan dari `LMSKE` [3] dengan modifikasi menggunakan FFmpeg untuk *scene detection* dan *histogram-based filtering*. Video diproses melalui proses **Scene-Based Extraction**: (1) *remuxing* untuk memperbaiki *timestamp* dengan flag `-fflags +genpts`, (2) *re-encoding* dengan `libx264` (parameter `-sc_threshold 40` untuk deteksi perubahan scene), dan (3) ekstraksi I-frame menggunakan filter `select='eq(pict_type,I)'`. Proses ini menghasilkan kandidat *keyframe* $\mathcal{F}_{\text{FFmpeg}} = \{(\mathbf{I}_i, t_i)\}_{i=1}^{N_{\text{raw}}}$ yang ter-*align* dengan transisi semantik, dengan $\mathbf{I}_i$ adalah *frame* ke-i; t_i adalah *timestamp* untuk *frame* bersesuaian; dan N_{raw} adalah banyaknya *keyframes*. Selanjutnya, kandidat *keyframe* dikonversi ke ruang warna HSV dengan **histogram** 3D ($8 \times 8 \times 8$ *bins*). *Frame* dipertahankan jika memiliki minimal 10 *non-zero bins* untuk mengeliminasi *frame uninformative*. Redundansi visual dihilangkan menggunakan *correlation coefficient* dengan threshold $\theta_{\text{sim}} = 0{,}8$, menghasilkan *final keyframe set* $\mathcal{K}$ yang non-redundan.

Visual Processing: Captioning dengan BLIP. *Keyframe* dari setiap video di-*caption* menggunakan model BLIP Large [7] dengan ukuran *batch* 32. Teks *caption* yang dihasilkan kemudian ditranslasikan ke dalam Bahasa Indonesia menggunakan model

`Helsinki-NLP OPUS-MT`.

Audio Processing. Audio track diekstrak dari video dan diproses menggunakan model OpenAI's Whisper `Large-v3-Turbo` [8] untuk menghasilkan transkrip bertimestamp. Model diimplementasikan via Hugging Face `transformers` pipeline dengan chunking strategy: `chunk_length_s` = 30 dan `stride_length_s` = 5 untuk menangani audio panjang dengan overlapping windows. Output berupa segmen-segmen transkrip $\{T_i\}_{i=1}^{N}$ yang menyediakan informasi temporal presisi untuk *alignment* dengan *visual keyframes*, dengan $T_i = \left(t_{\text{start}}^{(i)}, t_{\text{end}}^{(i)}, \texttt{text}_i\right)$; t_{start}^{i} adalah *timestamp* awal transkrip; $t_{\text{end}}^{(i)}$ adalah *timestamp* akhir transkrip; dan $\texttt{text}_i$ adalah isi transkrip.

II.3 Content Analysis

II.3.1 Semantic Chaptering

Transkrip audio dan *caption keyframe* digabungkan untuk segmentasi semantik menggunakan `Gemini 2.0 Flash` dengan *constraint*: (1) *chapter* pertama dimulai pada $t = 0.0$s, (2) durasi minimum 5 detik tanpa *overlap*, (3) *chapter* terakhir berakhir pada timestamp akhir video, dan (4) judul maksimal 3 kata (25 karakter). Output berformat JSON:

$$\texttt{chapter}_j = \left(c_j, t_{\text{start}}^{(j)}, t_{\text{end}}^{(j)}, \texttt{title}_j\right),$$

dan menghasilkan 3-7 *chapter* per video yang koheren secara semantik. Catat bahwa c_j representasi abstrak untuk sebuah konten ke-i dan $\texttt{title}_j$ adalah judul yang bersesuaian.

II.3.2 Chapter Summarization

Setiap *chapter* diringkas menggunakan `Gemini 2.0 Flash` dengan memprioritaskan transkrip sebagai sumber primer dan *caption* sebagai konteks visual, menghasilkan ringkasan 3 – 5 kalimat dalam Bahasa Indonesia.

II.3.3 Taxonomy Mapping dan Klasifikasi

Penelitian ini menerapkan *direct taxonomy mapping* berbasis *semantic similarity* terhadap taksonomi IAB Content Taxonomy 3.1 [9], menggantikan *unsupervised clustering* ala BERTopic [4]. Untuk representasi semantik kami menggunakan `intfloat/`

`multilingual-e5-large`[5] dengan skema *asymmetric prompts* seperti praktik umum pada retrieval: `"query:"` untuk ringkasan chapter dan `"passage:"` untuk label taksonomi. *Seluruh* embedding (chapter, konteks, dan label taksonomi) di-L_2-normalisasi agar dot product ekuivalen dengan *cosine similarity*.

Embedding dan Context Pooling. Untuk menjaga konteks temporal lintas *chapter*, digunakan `exponential-decay pooling`. Asumsikan seluruh *embedding* (ringkasan dan taksonomi) telah dinormalisasi L_2 terlebih dahulu, yaitu:

$$\mathbf{c}_1 = \texttt{norm}\left(\mathbf{e}^{(1)}_{\text{summary}}\right),$$

dengan $\mathbf{c}_1$ adalah konteks awal; $\texttt{norm}(.)$ adalah fungsi untuk normalisasi L_2. Kemudian, konteks untuk chapter ke-j, yang dinyatakan dengan $\mathbf{c}_j$, dapat dihitung dengan:

$$\mathbf{c}_j = \texttt{norm}\left((1-\alpha)\,\mathbf{e}^{(j)}_{\text{summary}} + \alpha\,\mathbf{c}_{j-1}\right), \quad j \geq 2\,.$$

Pekerjaan ini menggunakan nilai $\alpha = 0{,}6$ sebagai bobot *carry-over* dari $\mathbf{c}_{j-1}$. Jika keselarasan rendah, yakni $\mathbf{c}_{j-1} \cdot \mathbf{e}^{(j)}_{\text{summary}} < 0{,}15$, maka pada langkah j nilai α diganti dengan $\alpha_{\text{eff}} = 0{,}18$.

Taxonomy Assignment. Skor kesesuaian dihitung antara konteks chapter ke-j, $\mathbf{c}_{j-1}$, dengan embedding sebuah taksonomi dihitung menggunakan *cosine similarity*:

$$\texttt{sim}(j,k) = \frac{\mathbf{c}_j \cdot \mathbf{e}^{(k)}_{\text{taxo}}}{||\mathbf{c}_j||\,||\mathbf{e}^{(k)}_{\text{taxo}}||} \in [-1,1]\,,$$

dengan $\mathbf{e}^{(k)}_{\text{taxo}}$ adalah embedding label taksonomi-k yang telah dinormalisasi. Label *chapter* yang dipilih adalah *top-3* label dengan $\texttt{sim}(j,k) \geq 0{,}70$. Sedangkan, label *video* diperoleh dengan mengagregasi skor lintas *chapter* (mis. skema *sum-of-similarity* tak-negatif) dan memilih label dengan skor agregat tertinggi.

II.4 Video-Level Processing

II.4.1 Video-Level Summarization

Tiga strategi digunakan untuk menghasilkan ringkasan video [6], yaitu *stuffing*, *map-reduce*, dan *refine*.

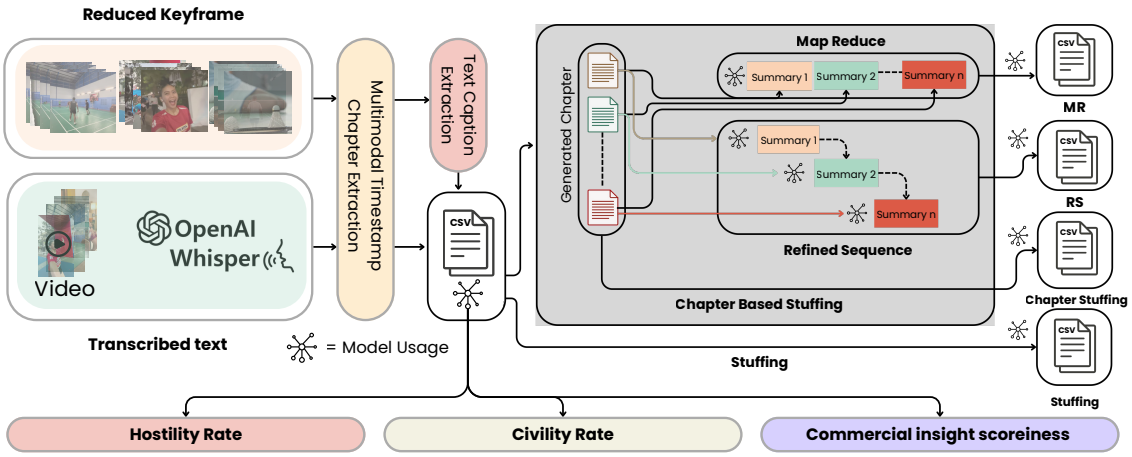


Figure 2. Proses ekstraksi pengetahuan multi-aspek dari kumpulan video.

Stuffing. Transkrip audio mentah dan informasi semantik visual digabungkan tanpa *preprocessing* dan diproses dalam satu *LLM call*:

$$\texttt{summary}(\text{video}) = \texttt{LLM}\left(\texttt{transcript}(\text{video}) \oplus \texttt{visual-sem}(\text{video})\right) ,$$

dengan $\texttt{LLM}(.)$ adalah sebuah *LLM call*; $\oplus$ adalah operator *concat* string deskripsi; $\texttt{transcript}(.)$ adalah fungsi ekstraksi transkripsi dari video; dan $\texttt{visual-sem}(.)$ adalah fungsi yang menghasilkan deskripsi semantik visual dari *frame-frame* video. Varian *chapter-based stuffing* menambahkan metadata temporal $(t^{(j)}_{\text{start}}, t^{(j)}_{\text{end}}, \texttt{title}_j)$ per segmen untuk konteks urutan kejadian yang lebih eksplisit.

MapReduce. Strategi ini memanfaatkan ringkasan *chapter* $\{s_j\}_{j=1}^{N_{\text{chapters}}}$ yang telah dihasilkan pada tahap *Chapter Summarization* (Subbab II.3, II.3.2). Ringkasan video dibentuk melalui fase *reduce* yang mengagregasi seluruh ringkasan *chapter*:

$$\texttt{summary}(\text{video}) = \texttt{LLM}_{\text{reduce}}\left(\bigoplus_{j=1}^{N_{\text{chapters}}} s_j\right) ,$$

dengan s_j adalah ringkasan *chapter* ke-j, $\texttt{LLM}_{\text{reduce}}(.)$ adalah *LLM call* untuk fase reduksi, dan $\bigoplus$ adalah operator konkatenasi berurutan. Strategi ini mengatasi *context length limitation* dengan memanfaatkan struktur *chapter* yang telah ada, tanpa memerlukan pemrosesan ulang konten mentah.

Refine. Ringkasan video dibangun secara iteratif dengan memperbarui ringkasan semen-

tara menggunakan informasi *chapter* baru pada setiap langkah:

$$\texttt{summary}_j^{\text{refine}} = \texttt{LLM}\left(\texttt{summary}_{j-1}^{\text{refine}} \oplus \texttt{summary}(c_j)\right),$$

dengan kondisi awal $\texttt{summary}_0^{\text{refine}} = \texttt{summary}(c_1)$ dan ringkasan final $\texttt{summary}(\text{video}) = \texttt{summary}_{N_{\text{chapters}}}^{\text{refine}}$. Strategi ini mempertahankan koherensi temporal dan konteks kumulatif, namun memerlukan N_{chapters} *sequential LLM calls*.

Struktur Output Ringkasan. Ketiga strategi menghasilkan ringkasan terstruktur dua paragraf dengan pedoman ketat: **Paragraf 1** menyajikan ringkasan komprehensif seluruh poin kunci sebagai daftar naratif yang padat dengan kalimat pengantar eksplisit (misal: "Video membahas..."), memberikan gambaran lengkap isi video. **Paragraf 2** diawali frasa penanda temporal ("Urutan yang disebut meliputi:" atau varian sejenis), diikuti elaborasi kronologis setiap poin sesuai kemunculan di video, dan diakhiri dengan penyebutan kesimpulan atau CTA jika ada. Ringkasan ditulis dalam gaya deskriptif pihak ketiga, *strictly grounded* pada transkrip dan *caption* tanpa penambahan informasi eksternal, dengan prioritas ekstraksi detail numerik dan spesifik secara harfiah daripada generalisasi.

II.4.2 Multi-Dimensional Content Scoring

Setiap video dinilai pada tiga dimensi menggunakan Gemini 2.0 Flash dengan input transkrip audio dan deskripsi semantik visual dari *keyframes*. Model menghasilkan skor $0 - 100$ untuk:

1. **Civility Score** berdasarkan kesopanan dan konstruktivitas nada komunikasi ($0 - 30$: kasar; $31 - 70$: netral; $71 - 100$: sopan);
2. **Hostility Score** berdasarkan tingkat agresi dan konfrontasi ($0 - 30$: ramah; $31 - 70$: asertif; $71 - 100$: agresif);
3. **Commercial Insight Score** berdasarkan identifikasi nilai bisnis, produk, dan *call-to-action* ($0 - 25$: tidak komersial; $26 - 75$: identifikasi produk; $76 - 100$: strategi komersial lengkap).

II.5 Evaluation Metrics

Evaluasi kualitas ringkasan dilakukan menggunakan dua pendekatan komplementer: metrik otomatis berbasis *n-gram overlap* dan penilaian kualitatif menggunakan *LLM-as-a-Judge*.

II.5.1 Automatic Metrics

Ringkasan dievaluasi terhadap *ground truth* manual menggunakan metrik standar: (1) **BLEU** mengukur presisi *n-gram* dengan *brevity penalty*, sensitif terhadap urutan kata yang tepat. (2) **ROUGE** mengukur *recall*-oriented overlap: ROUGE-1 (unigram), ROUGE-2 (bigram), dan ROUGE-L (*longest common subsequence*) untuk menangkap kesamaan struktural. (3) **METEOR** memperluas evaluasi dengan mempertimbangkan sinonim, stemming, dan *word order*, memberikan korelasi lebih tinggi dengan penilaian manusia dibanding BLEU.

II.5.2 LLM-as-a-Judge Evaluation

Untuk mengatasi keterbatasan metrik otomatis yang tidak menangkap aspek semantik mendalam, evaluasi kualitatif menggunakan `Gemini 2.0 Flash` sebagai evaluator. Metode ini dipilih karena korelasi tinggi dengan penilaian manusia (Spearman $\rho = 0.514$) dan skalabilitas untuk evaluasi berskala besar [10]. Setiap ringkasan dievaluasi pada empat dimensi dengan skala $1 - 100$:

1. **Akurasi (*Faithfulness*).** Mengukur keakuratan faktual ringkasan terhadap konten video tanpa penambahan informasi palsu;
2. **Kedalaman (*Depth*).** Mengukur kelengkapan cakupan poin-poin penting dari video. Skor tinggi (≥ 80) mengindikasikan ringkasan menangkap $> 90\%$ informasi kritis;
3. **Keringkasan (*Conciseness*).** Mengukur efisiensi penyampaian informasi tanpa redundansi berlebihan;
4. **Koherensi (*Coherence*).** Mengukur struktur logis dan keterbacaan ringkasan, mencakup transisi antar-ide dan alur naratif.

Untuk setiap metode *summarization*, LLM evaluator menerima transkrip audio, *visual captions*, dan ringkasan yang dihasilkan. Skor agregat dihitung sebagai rata-rata aritmatika dari akurasi, kedalaman, keringkasan, dan koherensi.

III HASIL DAN PEMBAHASAN

Bab ini mengkaji dua aspek penting: Evaluasi Kualitas Ringkasan Video (Subbab III.1), yang mengukur performa peringkasan, dan *Content Scoring* serta *Taxonomy Mapping* (Subbab III.2), yang menyediakan analisis konten mendalam.

III.1 Evaluasi Kualitas Ringkasan Video

III.1.1 Evaluasi dengan LLM-as-a-Judge

Tabel 2 menunjukkan perbandingan kinerja lima metode *summarization*. **MapReduce** (84,20) mengungguli metode lain dengan keseimbangan optimal pada kedalaman (91,00), keringkasan (77,50), dan koherensi (92,80).

Table 2. Perbandingan Metode Summarization dengan LLM-as-a-Judge

Metode	Akurasi	Kedalaman	Keringkasan	Koherensi	Rata-rata
Ground Truth	94,80	94,00	80,50	92,50	**90,45**
MapReduce	75,50	**91,00**	**77,50**	**92,80**	**84,20**
Chapter-based Stuffing	**97,50**	93,30	51,50	75,50	79,45
Stuffing	92.,0	94,70	49,10	78,00	78,53
Refine	72,00	85,80	54,50	73,00	71,33

Temuan utama: *Chapter-based Stuffing* mencapai akurasi tertinggi (97,50) namun dengan keringkasan rendah (51,50) akibat redundansi. *Refine* menunjukkan performa terendah (71,33) karena *information dilution* pada pemrosesan sekuensial.

III.1.2 Evaluasi dengan Metrik Otomatis

Tabel 3 menunjukkan pola berbeda: *Stuffing* unggul pada seluruh metrik *n-gram overlap* (ROUGE-1: 0,5005, METEOR: 0,4150) namun berada di peringkat ketiga pada LLM-as-a-Judge (78,53), mengonfirmasi keterbatasan metrik otomatis yang tidak menangkap aspek semantik seperti koherensi naratif [10].

III.2 Content Scoring dan Taxonomy Mapping

Sistem melakukan analisis multi-dimensi pada 1000 video dengan hasil distribusi skor yang mencerminkan karakteristik konten platform:

Table 3. Perbandingan Metode dengan Metrik Otomatis

Metode	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Stuffing	**0,1543**	**0,5005**	**0,2274**	**0,3556**	**0,4150**
Chapter-based Stuffing	0,1487	0,4949	0,2262	0,3484	0,4088
MapReduce	0,1054	0,4675	0,1701	0,2932	0,3620
Refine	0,0872	0,4394	0,1478	0,2503	0,3159

III.2.1 Analisis Content Scoring

Tabel 4 menampilkan statistik deskriptif untuk tiga dimensi penilaian konten.

Table 4. Distribusi Skor Konten Video

Metrik	Civility	Hostility	Commercial Insight
Mean ± SD	65,4 ± 6,2	11,6 ± 10,4	60,1 ± 20,3
Median	65,0	10,0	70,0
Min - Max	35 - 95	5 - 85	0 - 90

Temuan utama: (1) *Civility scores* menunjukkan distribusi terkonsentrasi dengan standar deviasi rendah (σ = 6,2), dengan mayoritas konten (88,6%) berada pada kategori netral-sopan (skor 60 – 80), mengindikasikan kualitas komunikasi yang konsisten terjaga. (2) *Hostility scores* memiliki median sangat rendah (10,0) dengan nilai maksimum 85, menunjukkan sebagian besar konten bersifat non-agresif dengan sebaran outlier hostile yang terdeteksi pada < 5% video. (3) *Commercial Insight* menunjukkan variance tertinggi (σ = 20,3) dengan median (70,0) lebih tinggi dari mean (60,1), mengindikasikan distribusi *left-skewed* dengan dominasi konten bermuatan komersial tinggi namun tetap terdapat subset signifikan konten edukatif murni (skor < 30).

III.2.2 Taxonomy Mapping

Taxonomy mapping dilakukan terhadap IAB Content Taxonomy 3.1 dengan hasil distribusi kategori tingkat atas (Tier-1) yang diringkas pada Tabel 5. Delapan kategori teratas mencakup ~85,44% dari seluruh penugasan (lihat persentase per baris). Cakupan kedalaman label ditunjukkan pada Tabel 6.

Komposisi Tier-1 didominasi *Style & Fashion* (23,31%), *Automotive* (15,56%), dan *Sports* (14,27%). Dari sisi kedalaman, mayoritas penugasan berhenti di Tier-2 (48,97%) dan Tier-3 (38,51%), dengan sebagian kecil mencapai Tier-4 (3,12%).

Table 5. Distribusi kategori Tier-1 (Top-8)

Kategori	Jumlah	Persentase
Style & Fashion	2.693	23,31%
Automotive	1.798	15,56%
Sports	1.649	14,27%
Genres	1.220	10,56%
Healthy Living	1.177	10,19%
Technology & Computing	714	6,18%
Travel	359	3,11%
Medical Health	261	2,26%

Table 6. Cakupan kedalaman pemetaan taksonomi IAB

Kedalaman	Jumlah	Persentase
Hanya Tier-1	1.086	9,40%
Hanya sampai Tier-2	5.658	48,97%
Hanya sampai Tier-3	4.450	38,51%
Sampai Tier-4	361	3,12%

IV PENUTUP

IV.1 Kesimpulan

Penelitian ini mengusulkan *chaptering-based multi-modal pipeline* yang efisien untuk *video understanding* pada video pendek. Kontribusi utama meliputi: (i) segmentasi semantik berbasis transkrip dan *keyframe captions*, (ii) *taxonomy mapping* langsung ke IAB 3.1 menggunakan `multilingual-e5-large` dengan *asymmetric prompts* dan *exponential-decay pooling* untuk konteks temporal, dan (iii) evaluasi tiga strategi peringkasan dengan metrik otomatis dan *LLM-as-a-Judge*. Pipeline mengadopsi arsitektur hibrid yang menggabungkan pemrosesan lokal dengan *cloud-based LLM* (`Gemini 2.0 Flash API`), memungkinkan implementasi yang lebih aksesibel dibanding *full local deployment* dengan Large VLM.

Evaluasi menunjukkan **MapReduce** memberikan keseimbangan terbaik (skor 84,20) pada koherensi dan keringkasan, sementara **Stuffing** mendominasi metrik *n-gram overlap* (ROUGE-1: 0,5005) namun lemah pada aspek semantik [10]. *Taxonomy mapping* pada 1000 video menghasilkan distribusi didominasi *Style & Fashion* (23,31%), *Automotive* (15,56%), dan *Sports* (14,27%), dengan mayoritas label mencapai Tier-2 (48,97%) dan Tier-3 (38,51%), menunjukkan kemampuan klasifikasi granular. Tujuan merancang

pipeline efisien untuk *video understanding* komprehensif tercapai dengan kebaruan pada integrasi *chaptering*, *context pooling* temporal, dan klasifikasi taksonomi standar industri [9, 5].

IV.2 Rekomendasi

(1) **Optimasi Taxonomy Mapping**: Implementasi kalibrasi ambang pada *dev set*, skema agregasi berbobot (*confidence-weighted voting*), indeks ANN (FAISS/ScaNN) untuk efisiensi pencarian, dan peningkatan penetrasi Tier-4 (saat ini 3,12%). (2) **Peningkatan Kualitas Summarization**: Eksplorasi strategi hibrid MapReduce dengan retensi konteks global, serta peningkatan kualitas ASR dan *visual captioning* untuk mengurangi propagasi error. (3) **Eksplorasi Model Lokal**: Evaluasi *small open-source LLM* (`Qwen2.5-7B`, `Llama-3-8B`) sebagai pengganti Gemini API untuk *full local deployment* dengan analisis *performance-resource trade-off*. (4) **Validasi dan Deployment**: Perluasan dataset beranotasi (> 50 video) dengan evaluasi manusia, serta implementasi kuantisasi model dan *batching* untuk skenario *production*.

REFERENSI

[1] Shuai Bai dkk. "Qwen2.5-VL Technical Report". In: *arXiv preprint* (2025). URL: https://arxiv.org/abs/2502.13923.

[2] Lucas Ventura dkk. "Chapter-Llama: Efficient Chaptering in Hour-Long Videos with LLMs". In: *arXiv preprint arXiv:2504.00072* (2025). URL: https://arxiv.org/abs/2504.00072.

[3] Kailong Tan dkk. "Large Model based Sequential Keyframe Extraction for Video Summarization". In: *arXiv preprint arXiv:2401.04962* (2024). URL: https://arxiv.org/abs/2401.04962.

[4] Maarten Grootendorst. "BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure". In: *arXiv preprint arXiv:2203.05794* (2022). URL: https://arxiv.org/abs/2203.05794.

[5] Liang Wang dkk. "Text Embeddings by Weakly-supervised Contrastive Pre-training". In: *arXiv preprint arXiv:2212.03533* (2024).

[6] LangChain. *Summarization Strategies Documentation*. 2024. URL: https://python.langchain.com/docs/.

[7] Junnan Li dkk. “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. 2022, pp. 12888–12900. URL: https://arxiv.org/abs/2201.12086.

[8] Alec Radford dkk. “Robust Speech Recognition via Large-Scale Weak Supervision”. In: *arXiv preprint arXiv:2212.04356* (2023). URL: https://arxiv.org/abs/2212.04356.

[9] IAB Tech Lab. *Content Taxonomy 3.1*. 2024. URL: https://github.com/InteractiveAdvertisingBureau/Taxonomies.

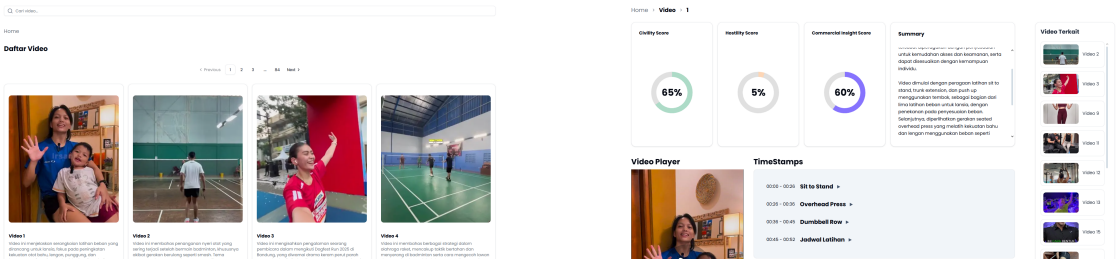
[10] Yang Liu dkk. “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023, pp. 2511–2522. URL: https://arxiv.org/abs/2303.16634.

V LAMPIRAN

V.1 Link Dashboard

Dashboard dapat diakses melalui: https://tim-suika.vercel.app/

V.2 Screenshot Dashboard



(a) (b)

Figure 3. Tampilan dashboard