🔍 STEP A1 — Dataset Inventory & Folder Integrity

1️⃣ Insight Detail per Sub-step / Output

📄 Output A1.1 — Video Folder Inventory

Fakta Penting (page 1–2)

- Terdapat **1000 file video** pada direktori:
  `/bdd100k/videos/train`.
- Seluruh file memiliki ekstensi **`.mov`**.
- Ukuran file video berada pada rentang **~5.34 MB hingga ~25.44 MB**.
- Mayoritas video berada di kisaran **18–20 MB**.
- Tidak ditemukan file non-video atau file tanpa ekstensi.

Insight

- Dataset video train **lengkap dan konsisten secara file-level**.
- Tidak ada indikasi file hilang, salah format, atau tercampur artefak non-video.

Ini menunjukkan

- Proses kurasi dataset BDD100K untuk split train dilakukan secara terkontrol.
- Video kemungkinan telah melalui proses preprocessing awal (durasi, resolusi, codec) yang relatif seragam.

📄 Output A1.2 — Video Size Statistics & Abnormal Check

Fakta Penting (page 3)

- Statistik ukuran video:
  - Mean: **~18.88 MB**
  - Median (50%): **~19.39 MB**
  - Min: **5.34 MB**
  - Max: **25.44 MB**
- Threshold file kecil didefinisikan `< 1 MB`.
- **Tidak ada** video yang berada di bawah threshold tersebut.
- Distribusi ukuran relatif sempit (std ~2.19 MB).

Insight

- Tidak ditemukan **file video rusak, kosong, atau terpotong**.
- Variasi ukuran masih dalam batas wajar untuk dataset video driving.

Ini menunjukkan

- Seluruh video dapat diasumsikan **readable dan usable** pada tahap decoding frame.
- Risiko kegagalan `cv2.VideoCapture` akibat file corruption sangat rendah.

Output A1.3 — Video Format Summary

Fakta Penting (page 4)

- Seluruh **1000 video menggunakan format `.mov`**.
- Tidak ada variasi format video lain (mis. `.mp4`, `.avi`).

Insight

- Format video **sepenuhnya homogen**.

Ini menunjukkan

- Pipeline preprocessing video dapat disederhanakan:
  - Tidak perlu branching logic berbasis format.
  - Decoder dan parameter IO dapat diasumsikan konsisten.

Output A1.4 — Label File Accessibility Check

Fakta Penting (page 5)

- `mot_labels.csv`:
  - Berhasil dibaca
  - Shape: **(2,890,846 rows × 13 columns)**
- `mot_labels.parquet`:
  - Berhasil dibaca
  - Shape: **(2,890,846 rows × 13 columns)**
- Tidak ada error IO atau schema mismatch awal.

Insight

- Kedua format label **valid secara teknis** dan **konsisten isinya**.

Ini menunjukkan

- Dataset menyediakan **dua representasi label yang ekuivalen**, sehingga pemilihan format dapat didasarkan pada:
    - performa IO
    - efisiensi memory
    - kemudahan join & groupby di tahap lanjut

Output A1.5 — Dataset Inventory Summary

Fakta Penting (page 6)

- Total video train: **1000**
- Total ukuran video: **~18.44 GB**
- Status label:
    - CSV: OK
    - Parquet: OK

Insight

- Dataset **siap diproses** dari sisi ketersediaan aset mentah.

Ini menunjukkan

- Tidak ada blocker teknis awal yang menghalangi masuk ke validasi kontrak label (STEP A2).

---

2 Insight Kesimpulan (Naratif — Mendalam)

STEP A1 mengonfirmasi bahwa dataset BDD100K pada level file dan aset mentah berada dalam kondisi sangat sehat. Seluruh video train tersedia lengkap, menggunakan satu format homogen (`.mov`), dengan distribusi ukuran yang stabil dan tanpa indikasi file rusak atau abnormal. Dari sisi label, baik versi CSV maupun Parquet dapat diakses dengan sukses dan menunjukkan konsistensi dimensi yang identik, menandakan tidak adanya masalah korupsi atau mismatch awal.

Dengan demikian, seluruh prasyarat teknis dasar—mulai dari keterbacaan video hingga ketersediaan label—telah terpenuhi. STEP ini berfungsi sebagai *gatekeeper* yang memastikan bahwa seluruh EDA lanjutan berdiri di atas dataset yang valid secara teknis, bukan sekadar “tersedia”.

3️⃣ Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

Tidak ditemukan masalah teknis fatal pada STEP A1.

Hipotesis risiko minor (bersifat potensial, bukan temuan):

- Variasi ukuran video (5–25 MB) mungkin mencerminkan perbedaan durasi atau bitrate, yang **perlu diperhatikan saat frame sampling**, tetapi **bukan isu integritas**.

4️⃣ Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Pipeline decoding video dapat diasumsikan:
  - satu format (`.mov`)
  - tanpa kebutuhan fallback untuk file rusak
- Label dapat dipilih salah satu format:
  - **Parquet** lebih cocok untuk operasi EDA lanjutan dan join besar
  - **CSV** tetap valid sebagai referensi atau interoperabilitas

Kaitan dengan Modeling

- Tidak ada constraint awal yang memaksa:
  - pengurangan dataset
  - filtering video
- Modeling dapat diasumsikan bekerja pada **1000 video utuh**, dengan skala data yang konsisten.

🔍 STEP A1 — Dataset Inventory & Folder Integrity

1️⃣ Insight Detail per Sub-step / Output

🔍 Output A1.1 — Video Folder Inventory

Fakta Penting (page 1–2)

- Terdapat **1000 file video** pada direktori: `/bdd100k/videos/train`.
- Seluruh file memiliki ekstensi **`.mov`**.
- Ukuran file video berada pada rentang **~5.34 MB hingga ~25.44 MB**.
- Mayoritas video berada di kisaran **18–20 MB**.
- Tidak ditemukan file non-video atau file tanpa ekstensi.

Insight

- Dataset video train **lengkap dan konsisten secara file-level**.
- Tidak ada indikasi file hilang, salah format, atau tercampur artefak non-video.

Ini menunjukkan

- Proses kurasi dataset BDD100K untuk split train dilakukan secara terkontrol.
- Video kemungkinan telah melalui proses preprocessing awal (durasi, resolusi, codec) yang relatif seragam.

🔍 Output A1.2 — Video Size Statistics & Abnormal Check

Fakta Penting (page 3)

- Statistik ukuran video:
  - Mean: **~18.88 MB**
  - Median (50%): **~19.39 MB**
  - Min: **5.34 MB**
  - Max: **25.44 MB**
- Threshold file kecil didefinisikan `< 1 MB`.
- **Tidak ada** video yang berada di bawah threshold tersebut.
- Distribusi ukuran relatif sempit (std ~2.19 MB).

Insight

- Tidak ditemukan **file video rusak, kosong, atau terpotong**.
- Variasi ukuran masih dalam batas wajar untuk dataset video driving.

Ini menunjukkan

- Seluruh video dapat diasumsikan **readable dan usable** pada tahap decoding frame.
- Risiko kegagalan `cv2.VideoCapture` akibat file corruption sangat rendah.

Output A1.3 — Video Format Summary

Fakta Penting (page 4)

- Seluruh **1000 video menggunakan format `.mov`**.
- Tidak ada variasi format video lain (mis. `.mp4`, `.avi`).

Insight

- Format video **sepenuhnya homogen**.

Ini menunjukkan

- Pipeline preprocessing video dapat disederhanakan:
  - Tidak perlu branching logic berbasis format.
  - Decoder dan parameter IO dapat diasumsikan konsisten.

Output A1.4 — Label File Accessibility Check

Fakta Penting (page 5)

- `mot_labels.csv`:
  - Berhasil dibaca
  - Shape: **(2,890,846 rows × 13 columns)**
- `mot_labels.parquet`:
  - Berhasil dibaca
  - Shape: **(2,890,846 rows × 13 columns)**
- Tidak ada error IO atau schema mismatch awal.

Insight

- Kedua format label **valid secara teknis** dan **konsisten isinya**.

Ini menunjukkan

- Dataset menyediakan **dua representasi label yang ekuivalen**, sehingga pemilihan format dapat didasarkan pada:
    - performa IO
    - efisiensi memory
    - kemudahan join & groupby di tahap lanjut

Output A1.5 — Dataset Inventory Summary

Fakta Penting (page 6)

- Total video train: **1000**
- Total ukuran video: **~18.44 GB**
- Status label:
    - CSV: OK
    - Parquet: OK

Insight

- Dataset **siap diproses** dari sisi ketersediaan aset mentah.

Ini menunjukkan

- Tidak ada blocker teknis awal yang menghalangi masuk ke validasi kontrak label (STEP A2).

---

2 Insight Kesimpulan (Naratif — Mendalam)

2 Insight Kesimpulan (Naratif — Mendalam, Terhubung)

Sub-step A2.1
Proses awal memverifikasi bahwa label tersedia dalam dua format (CSV dan Parquet) dengan struktur yang sepenuhnya identik, baik dari sisi jumlah baris, kolom, maupun urutan kolom. Hal ini memastikan bahwa analisis lanjutan dapat difokuskan pada satu representasi label tanpa kekhawatiran adanya perbedaan semantik antar format, sekaligus menegaskan bahwa integritas label terjaga sejak tahap penyimpanan.

Sub-step A2.2
Setelah memastikan keterbacaan label, inspeksi schema mengungkap bahwa meskipun struktur kolom konsisten, tidak semua baris merepresentasikan anotasi objek yang lengkap. Kehadiran missing value pada kolom `id` dan `box2d.*` mengindikasikan bahwa label mencampur antara

baris anotasi objek dan baris non-objek atau metadata, yang secara implisit membentuk kontrak bahwa tidak semua baris layak langsung digunakan untuk training deteksi.

Sub-step A2.3
Dalam konteks temporal, kolom `frameIndex` terkonfirmasi menggunakan skema zero-based dengan rentang frame yang relatif terbatas. Temuan ini melengkapi pemahaman sebelumnya dengan menegaskan bahwa anotasi terikat kuat pada struktur video-frame, sehingga asumsi temporal (urutan frame dan kontinuitas waktu) dapat dibangun secara eksplisit pada tahap tracking dan frame sampling.

Sub-step A2.4
Pada dimensi semantik, eksplorasi kolom `category` memperlihatkan distribusi kelas yang sangat timpang dengan dominasi objek kendaraan, khususnya mobil. Ketimpangan ini bukan sekadar statistik distribusi, tetapi menjadi sifat inheren dataset yang harus dipertimbangkan sejak awal, karena ia akan berinteraksi langsung dengan keputusan filtering baris valid, desain evaluasi, dan strategi pembelajaran model.

Sub-step A2.5
Ketika beralih ke aspek identitas objek, analisis kolom identifier menunjukkan bahwa tidak ada satu kolom pun yang dapat berfungsi sebagai primary key tunggal. Track `id` tidak unik secara global dan hanya bermakna dalam konteks video tertentu, sehingga kontrak identitas label bersifat komposit. Temuan ini mengikat kembali hasil sebelumnya: integritas temporal dan semantik hanya dapat dijaga jika label diperlakukan sebagai kombinasi identifier, bukan entitas baris mandiri.

Sub-step A2.6
Pemeriksaan bounding box kemudian memperjelas representasi spasial anotasi. Nilai `box2d.x1, x2, y1, y2` berada pada skala ratusan hingga ribuan, menegaskan bahwa format bbox adalah **xyxy dengan koordinat pixel absolut**, bukan normalized. Keputusan ini melengkapi kontrak label dengan menetapkan aturan geometri yang eksplisit dan menghindari ambiguitas interpretasi pada tahap preprocessing maupun augmentasi.

Sub-step A2.7
Akhirnya, schema card menyatukan seluruh temuan sebelumnya menjadi kontrak label yang eksplisit: label BDD100K menggabungkan anotasi objek inti (bbox, category, track id) dengan atribut kontekstual tambahan (crowd, occluded, truncated) serta metadata video. Dengan pemahaman ini, label tidak lagi dipandang sebagai tabel datar, melainkan sebagai struktur anotasi berlapis yang memerlukan pemisahan peran kolom secara sadar sebelum masuk ke preprocessing dan modeling

---

3️⃣ Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

Tidak ditemukan masalah teknis fatal pada STEP A1.

Hipotesis risiko minor (bersifat potensial, bukan temuan):

- Variasi ukuran video (5–25 MB) mungkin mencerminkan perbedaan durasi atau bitrate, yang **perlu diperhatikan saat frame sampling**, tetapi **bukan isu integritas**.

4️⃣ Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Pipeline decoding video dapat diasumsikan:
  - satu format (`.mov`)
  - tanpa kebutuhan fallback untuk file rusak
- Label dapat dipilih salah satu format:
  - **Parquet** lebih cocok untuk operasi EDA lanjutan dan join besar
  - **CSV** tetap valid sebagai referensi atau interoperabilitas

Kaitan dengan Modeling

- Tidak ada constraint awal yang memaksa:
  - pengurangan dataset
  - filtering video
- Modeling dapat diasumsikan bekerja pada **1000 video utuh**, dengan skala data yang konsisten.

? STEP A3 — Video ↔ Label Joinability Check

---

1 Insight Detail per Sub-step / Output

? Output A3.1 — Video File List

Fakta Penting

- Direktori video train berisi **1000 file video**.
- Seluruh video berada pada path yang konsisten.
- Daftar video dijadikan *ground truth* keberadaan video fisik.

Insight

- Set video train dapat didefinisikan secara deterministik tanpa ambiguitas file system.

Ini menunjukkan

- Setiap proses join label harus mengacu pada daftar 1000 video ini sebagai referensi final.

Resiko

- –

? Output A3.2 — Label Video Identifiers

Fakta Penting

- Kolom `videoName` pada label memiliki **1400 nilai unik** (baik CSV maupun Parquet).
- Jumlah ini **lebih besar** dari jumlah video fisik (1000).

Insight

- Tidak semua identifier video pada label memiliki pasangan video fisik.

Ini menunjukkan

- Label mencakup anotasi untuk video di luar subset train yang tersedia.

Resiko

- Label dapat mengandung anotasi *out-of-scope* jika tidak difilter berdasarkan video fisik.

□ Output A3.3 — Video Identifier Normalization

Fakta Penting

- Normalisasi dilakukan dengan menambahkan ekstensi `.mov` jika belum ada.
- Jumlah identifier ternormalisasi tetap **1400** (CSV & Parquet).

Insight

- Perbedaan join bukan disebabkan oleh format penamaan, tetapi oleh cakupan dataset.

Ini menunjukkan

- Mismatch video ↔ label bersifat struktural (subset selection), bukan kesalahan penulisan nama.

Resiko

- –

□ Output A3.4 — Coverage Report (CSV)

Fakta Penting

- Dari 1000 video:
  - **961** video memiliki label.
  - **39** video tidak memiliki label.
- Terdapat **439 orphan label videos** (label tanpa video fisik).

Insight

- Joinability antara video dan label **tidak sempurna**.

Ini menunjukkan

- Dataset train video merupakan **subset** dari keseluruhan label BDD100K.

Resiko

- Jika orphan label tidak difilter, dapat terjadi error saat loading video atau silent data leakage.

Output A3.5 — CSV vs Parquet Coverage Comparison

Fakta Penting

- Coverage CSV dan Parquet **identik**:
  - 961 video berlabel
  - 39 video tanpa label
  - 439 orphan label

Insight

- Tidak ada perbedaan joinability antara CSV dan Parquet.

Ini menunjukkan

- Pemilihan sumber label tidak memengaruhi integritas join video ↔ label.

Resiko

- –

Output A3.6 — Mismatch Listing

Fakta Penting

- Daftar eksplisit tersedia untuk:
  - Video fisik tanpa label
  - Label tanpa video fisik
- Daftar mismatch konsisten antara CSV dan Parquet.

Insight

- Mismatch bersifat deterministik dan dapat ditangani dengan filtering eksplisit.

Ini menunjukkan

- Tidak ada ambiguity acak; seluruh mismatch dapat direproduksi dan diaudit.

Resiko

- Jika video tanpa label tidak ditangani, dapat memengaruhi split atau evaluasi.

Output A3.7 — Joinability Decision

Fakta Penting

- CSV dan Parquet menunjukkan coverage yang sama.
- Rekomendasi: **gunakan Parquet** untuk tahap lanjut.

Insight

- Sumber label dapat diputuskan berdasarkan efisiensi teknis, bukan isi data.

Ini menunjukkan

- Parquet aman dijadikan **single source of truth** untuk EDA dan preprocessing berikutnya.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step A3.1
Daftar video fisik pada split train berhasil ditetapkan secara eksplisit sebagai himpunan referensi utama, sehingga seluruh proses join label selanjutnya memiliki landasan deterministik yang jelas dan tidak bergantung pada asumsi implisit.

Sub-step A3.2
Ketika identifier video dari label diekstraksi, segera terlihat bahwa cakupan label lebih luas daripada video train yang tersedia. Hal ini mengindikasikan bahwa label BDD100K bersifat global, sementara dataset video yang digunakan merupakan subset terkurasi.

Sub-step A3.3
Upaya normalisasi penamaan menegaskan bahwa perbedaan cakupan bukan disebabkan oleh inkonsistensi format nama, melainkan oleh perbedaan ruang lingkup dataset. Dengan demikian, mismatch yang muncul bersifat struktural, bukan kesalahan teknis sederhana.

Sub-step A3.4
Analisis coverage memperlihatkan bahwa sebagian besar video train telah berlabel, namun tetap terdapat video tanpa label serta label tanpa video. Temuan ini mengonfirmasi bahwa join video ↔ label memerlukan filtering eksplisit agar pipeline tidak memproses data di luar ruang lingkup.

Sub-step A3.5
Perbandingan antara CSV dan Parquet menunjukkan kesetaraan penuh dalam hasil joinability, menghilangkan kemungkinan adanya perbedaan isi atau kehilangan data antar format label.

Sub-step A3.6
Daftar mismatch yang terdefinisi dengan jelas memastikan bahwa seluruh ketidaksesuaian dapat ditangani secara deterministik. Tidak ada kasus abu-abu yang memerlukan inspeksi manual tambahan.

Sub-step A3.7
Keputusan akhir menetapkan Parquet sebagai sumber label utama, mengikat seluruh temuan sebelumnya ke satu pilihan teknis yang konsisten dan efisien untuk EDA serta preprocessing lanjutan.

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Dataset video train merupakan subset dari keseluruhan label BDD100K.
- Terdapat video tanpa label dan label tanpa video yang harus difilter.
- Masalah ini bersifat struktural dan dapat ditangani dengan aturan join yang eksplisit.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Label **wajib difilter** berdasarkan daftar video fisik train.
- Orphan label harus di-drop sebelum frame extraction.
- Video tanpa label perlu diputuskan: di-drop atau dipisahkan (mis. untuk unsupervised use).

Kaitan dengan Modeling

- Training dan evaluasi hanya boleh menggunakan video yang **joinable**.
- Tidak boleh ada asumsi bahwa seluruh label BDD100K relevan dengan split train ini.
- Tracking dan evaluation protocol harus berbasis subset video yang telah tervalidasi.

✅ STEP A4 — Early Assumption Validation (Fail Fast)

1️⃣ Insight Detail per Sub-step / Output

✅ Output A4.1 — Label Load & Basic Filtering

Fakta Penting

- Setelah filtering baris tanpa bbox, id, frameIndex, dan videoName, tersisa **2,886,916 baris label**.
- Filtering hanya menghilangkan baris yang tidak memenuhi kontrak anotasi objek minimum.

Insight

- Mayoritas besar label merepresentasikan anotasi objek yang valid dan siap dianalisis lebih lanjut.

Ini menunjukkan

- Dataset label memiliki densitas anotasi tinggi, dan baris non-objek bukan komponen dominan.

Resiko

- –

✅ Output A4.2 — One Video = One Sequence

Fakta Penting

- Untuk setiap `videoName`, frameIndex selalu dimulai dari **0**.
- Nilai maksimum frameIndex per video berada di kisaran **201–202**.
- Jumlah frame unik per video konsisten dengan rentang tersebut.

Insight

- Satu video merepresentasikan **satu sequence temporal utuh**.

Ini menunjukkan

- Tidak ada indikasi bahwa satu video berisi beberapa sequence terpisah atau terfragmentasi.

Resiko

- –

Output A4.3 — Frame Continuity per Track

Fakta Penting

- Sebagian besar track memiliki **0 gap frame** (90.057 track).
- Namun terdapat track dengan gap frame >0, dengan distribusi ekor panjang hingga >20 gap.

Insight

- Tidak semua track bersifat kontinu secara temporal; sebagian objek mengalami **track fragmentation**.

Ini menunjukkan

- Tracking annotation tidak menjamin keberlanjutan objek di setiap frame berturut-turut, kemungkinan akibat occlusion, keluar-masuk frame, atau kebijakan anotasi.

Resiko

- Jika model atau preprocessing mengasumsikan track selalu kontinu, hasil tracking atau sequence modeling dapat bias.

Output A4.4 — Track ID Scope

Fakta Penting

- Total unique track id: **112.819**.
- **Tidak ada** track id yang muncul di lebih dari satu video.

Insight

- Track id bersifat **lokal per video**, bukan global.

Ini menunjukkan

- Identitas objek hanya bermakna dalam konteks satu video dan tidak boleh dibandingkan lintas video.

Resiko

- –

Output A4.5 — Duplicate Core Key Check

Fakta Penting

- Tidak ditemukan duplikasi pada kombinasi **(videoName, frameIndex, id)**.

Insight

- Kontrak identitas anotasi pada level frame-track terjaga dengan baik.

Ini menunjukkan

- Setiap objek pada satu frame dalam satu video direpresentasikan secara unik.

Resiko

- –

Output A4.6 — Assumption Validation Summary

Fakta Penting

- Seluruh asumsi dasar telah diuji:
  - video = sequence
  - frame index terurut
  - track id lokal
  - tidak ada duplikasi core key

Insight

- Sebagian besar asumsi dasar **valid**, dengan satu pengecualian penting terkait kontinuitas track.

Ini menunjukkan

- Dataset relatif aman untuk pipeline standar MOT, dengan catatan khusus pada fragmentasi track.

Resiko

- Jika fragmentasi track diabaikan, analisis temporal atau tracking-based modeling dapat menghasilkan interpretasi keliru.

2️⃣ Insight Kesimpulan (Naratif — Terhubung)

Sub-step A4.1
Filtering awal memastikan bahwa hanya anotasi dengan kontrak minimum yang dianalisis, sekaligus menunjukkan bahwa sebagian besar label memang merepresentasikan objek valid dan bukan artefak metadata.

Sub-step A4.2
Dengan frame index yang konsisten dimulai dari nol dan membentuk rentang utuh per video, asumsi bahwa satu video merepresentasikan satu sequence temporal dapat diterima tanpa reservasi tambahan.

Sub-step A4.3
Namun, ketika masuk ke level track, terlihat bahwa kontinuitas temporal tidak selalu terjaga. Keberadaan gap pada sebagian track mengindikasikan bahwa anotasi bersifat realistis dan mencerminkan kondisi lapangan, bukan sequence objek ideal yang selalu muncul di setiap frame.

Sub-step A4.4
Validasi scope track id memperkuat pemahaman bahwa identitas objek sepenuhnya terikat pada video, sehingga tidak ada konsep objek global lintas video dalam dataset ini.

Sub-step A4.5
Tidak ditemukannya duplikasi pada core key menegaskan bahwa struktur label konsisten dan aman dari konflik identitas pada level frame-track.

Sub-step A4.6
Secara keseluruhan, STEP A4 berhasil memvalidasi asumsi-asumsi kritis pipeline MOT, sekaligus menandai satu batasan penting yang perlu diperhatikan pada tahap berikutnya, yaitu fragmentasi track.

3️⃣ Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Track tidak selalu kontinu secara temporal.
- Fragmentasi kemungkinan disebabkan oleh occlusion, objek keluar-masuk frame, atau kebijakan anotasi.
- Masalah ini bersifat **inheren pada dataset**, bukan kesalahan anotasi.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Preprocessing **tidak boleh mengasumsikan track kontinu**.
- Perlu mekanisme eksplisit untuk menangani gap frame (mis. interpolation, padding, atau pemutusan sequence).
- Sequence construction harus berbasis frameIndex aktual, bukan panjang track ideal.

Kaitan dengan Modeling

- Model tracking harus toleran terhadap fragmentasi track.
- Model temporal (mis. RNN/Transformer di atas track) perlu desain yang tidak bergantung pada kontinuitas penuh.
- Evaluasi tracking harus mempertimbangkan bahwa gap merupakan sifat data, bukan error model.

▯ STEP A5 — Video Decode-Level Sanity Check

1▯ Insight Detail per Sub-step / Output

▯ Output A5.1 — Video File List

Fakta Penting

- Terdapat **1000 file video** pada direktori `bdd100k/videos/train`.
- Seluruh video terdeteksi sesuai dengan inventory pada STEP A1.

Insight

- Set video yang diuji pada decode-level sanity **identik dengan set video train**.

Ini menunjukkan

- Tidak ada perbedaan atau kehilangan file antara inventory file-level dan proses decoding.

Resiko

- –

▯ Output A5.2 — Decode-Level Metadata Extraction

Fakta Penting

- **1000 video berhasil dibuka** menggunakan `cv2.VideoCapture`.
- **0 video gagal decode**.
- Metadata berhasil diekstraksi untuk seluruh video.

Insight

- Seluruh video **valid secara decode-level**.

Ini menunjukkan

- Risiko kegagalan frame extraction akibat file corruption atau codec issue sangat rendah.

Resiko

- –

Output A5.3 — Videos Failed to Decode

Fakta Penting

- Tidak ditemukan video yang gagal dibuka decoder.

Insight

- Tidak ada subset video bermasalah secara teknis.

Ini menunjukkan

- Dataset tidak memerlukan pengecualian khusus pada tahap decoding.

Resiko

- –

Output A5.4 — Metadata Distribution Summary

Fakta Penting

- Rata-rata jumlah frame: **~1.179 frame** per video.
- Median frame count: **~1.207 frame**.
- Frame count minimum: **313**, maksimum: **1.567**.
- FPS sangat konsisten di sekitar **30 FPS** (std ~0.11).
- Resolusi seluruh video **seragam: 1280 × 720**.

Insight

- Dataset video **sangat terstandardisasi** dari sisi durasi, FPS, dan resolusi.

Ini menunjukkan

- Pipeline preprocessing dapat diasumsikan bekerja pada satu konfigurasi resolusi dan framerate tanpa normalisasi tambahan.

Resiko

- –

Output A5.5 — Extreme Duration Check

Fakta Penting

- Tidak ada video dengan durasi sangat pendek (<30 frame).
- Terdapat **9 video** dengan durasi di atas persentil ke-99.
- Video terpanjang memiliki **1.567 frame**.

Insight

- Outlier durasi **sangat terbatas jumlahnya** dan masih berada dalam rentang wajar.

Ini menunjukkan

- Tidak ada video ekstrem yang berpotensi merusak strategi sampling atau batching secara signifikan.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step A5.1
Inventarisasi ulang pada tahap decode memastikan bahwa seluruh video yang tercatat pada file system benar-benar masuk dalam proses validasi, tanpa adanya perbedaan set antara tahap inventory dan decoding.

Sub-step A5.2
Keberhasilan membuka seluruh video menggunakan decoder menunjukkan bahwa dataset aman dari masalah teknis mendasar seperti file corrupt atau codec tidak didukung, sehingga proses ekstraksi frame dapat dilakukan secara langsung.

Sub-step A5.3
Ketiadaan video yang gagal decode menghilangkan kebutuhan akan mekanisme fallback atau pengecualian khusus pada pipeline preprocessing.

Sub-step A5.4
Distribusi metadata yang sangat sempit—baik dari sisi FPS maupun resolusi—menegaskan bahwa dataset telah distandardisasi, memungkinkan desain preprocessing yang sederhana dan efisien.

Sub-step A5.5
Meskipun terdapat sejumlah kecil video berdurasi lebih panjang, jumlahnya sangat terbatas dan tidak membentuk kelas outlier yang mengkhawatirkan, sehingga tidak memerlukan penanganan khusus pada tahap awal.

3️⃣ Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Tidak ditemukan masalah decode-level pada dataset video.
- Variasi durasi video masih berada dalam batas wajar dan tidak bersifat problematik.

4️⃣ Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Frame extraction dapat dilakukan tanpa pengecekan tambahan terkait corrupt file.
- Resolusi dan FPS dapat diasumsikan tetap (1280×720 @ ~30 FPS).
- Sampling strategy dapat menggunakan parameter global tanpa conditional logic.

Kaitan dengan Modeling

- Model tidak perlu adaptasi khusus untuk perbedaan resolusi atau FPS.
- Video berdurasi panjang dapat diperlakukan sama dengan video lain atau dipotong secara sistematis jika diperlukan.

▯ STEP A6 — FrameIndex ↔ Video Frame Count Consistency Check

1▯ Insight Detail per Sub-step / Output

▯ Output A6.1 — Video Metadata Load

Fakta Penting

- Metadata video berhasil dimuat untuk **1000 video** (frame_count sebagai ground truth).

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Referensi durasi video sudah lengkap untuk evaluasi konsistensi label-frame.

Ini menunjukkan

- Sisi “video” tidak menjadi bottleneck untuk validasi alignment; semua video punya frame_count.

Resiko

- –

▯ Output A6.2 — Max FrameIndex per Video

Fakta Penting

- Label menghasilkan **1400 video unik** setelah normalisasi `.mov`.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Label memang mencakup video di luar subset train yang tersedia (konsisten dengan A3).

Ini menunjukkan

- Normalisasi key videoName sudah bekerja, dan A6 hanya perlu mengevaluasi subset train (1000 video).

Resiko

- –

Output A6.3 — Label ↔ Video Join

Fakta Penting

- Join menghasilkan **1000 baris** (tepat jumlah video train).

  EDA_STEP_A6_FrameIndex_VideoFra…

- `NaN max_frameIndex_label` dipakai untuk menandai **NO_LABEL**.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Joinability pada level identifier video sudah deterministik dan konsisten untuk evaluasi status per-video.

Ini menunjukkan

- Masalah “semua NO_LABEL” yang terjadi sebelumnya sudah terselesaikan; A6 sekarang mengukur hal yang benar.

Resiko

- –

Output A6.4 — Alignment Status Summary

Fakta Penting

- Status video: **VALID = 961**, **NO_LABEL = 39**, **INVALID = 0**.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Untuk seluruh video yang berlabel, **tidak ada kasus frameIndex label yang melampaui durasi video**.

Ini menunjukkan

- Tidak ada “silent mismatch fatal” tipe: label menunjuk frame yang tidak eksis (untuk subset berlabel).

Resiko

- Untuk 39 video NO_LABEL, pipeline training supervised harus eksplisit memutuskan perlakuannya agar tidak masuk train secara tidak sengaja (bukan risiko alignment, tapi risiko dataset assembly).

Output A6.5 — Scatter: FrameIndex vs FrameCount

Fakta Penting

- Plot menunjukkan titik-titik **jauh di bawah garis batas $y \leq x-1$** (tidak ada yang menembus batas).

  EDA_STEP_A6_FrameIndex_VideoFra…

- Secara visual, `max_frameIndex_label` tampak berkisar sekitar **~100–~210**, sementara frame_count video berada di kisaran **~600–~1600**. (terlihat pada page plot)

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Label frameIndex berada jauh di bawah panjang video → label tampaknya hanya mencakup **subset awal frame** dari tiap video, bukan seluruh durasi video.

Ini menunjukkan

- Kontrak label kemungkinan “sequence/window labeling” (mis. label hanya untuk sekitar ~200 frame awal), bukan full-video annotation—namun tetap konsisten secara indeks.

Resiko

- Jika preprocessing/training mengasumsikan “seluruh frame video berlabel”, maka frame-frame setelah `max_frameIndex_label` bisa salah diperlakukan sebagai background/negative secara implisit (label-missing ≠ negative).

Output A6.6 — Alignment Status Distribution

Fakta Penting

- Distribusi hanya berisi **VALID dan NO_LABEL**; tidak ada INVALID.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Kondisi “aman dari INVALID” bersifat sistemik, bukan kebetulan sporadis.

Ini menunjukkan

- Secara global, dataset (subset berlabel) memenuhi kontrak alignment yang diperlukan untuk training.

Resiko

- –

Output A6.7 — Invalid Alignment Listing

Fakta Penting

- Tercatat eksplisit: **tidak ada video** dengan frameIndex label melebihi decoded frame count.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Tidak diperlukan mekanisme filtering karena INVALID (karena memang 0).

Ini menunjukkan

- Step fail-fast ini lulus: tidak ada kasus label menunjuk frame non-eksis.

Resiko

- –

Output A6.8 — Alignment Decision Summary

Fakta Penting

- Ringkasan keputusan: **VALID 961, INVALID 0, NO_LABEL 39** dan rule keputusan sesuai.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Keputusan A6 adalah **alignment OK untuk subset berlabel**.

Ini menunjukkan

- Dataset siap lanjut ke langkah berikutnya tanpa perbaikan alignment, cukup penanganan 39 video NO_LABEL sebagai kasus coverage.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step A6.1
Dengan frame_count tersedia untuk 1000 video, A6 punya ground truth yang kuat untuk memastikan label tidak menunjuk ke frame di luar durasi video.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.2
Perhitungan max frameIndex pada label sekaligus menegaskan bahwa label mencakup video yang lebih luas (1400), sehingga evaluasi harus berbasis subset train yang memang ada.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.3
Join yang menghasilkan tepat 1000 baris dan penggunaan NO_LABEL via NaN menunjukkan proses pemetaan video ↔ label sudah stabil dan dapat dipercaya untuk evaluasi status per video.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.4
Hasil status (961 VALID, 39 NO_LABEL, 0 INVALID) menutup risiko fatal “label melampaui durasi video” untuk subset yang berlabel, sambil menandai adanya gap coverage yang harus diperlakukan khusus untuk 39 video.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.5
Scatter memperlihatkan pola penting: walaupun alignment aman, label tampaknya hanya meng-cover sebagian awal video (max_frameIndex sekitar ~200) sementara videonya jauh lebih panjang, sehingga kontrak data lebih mirip window/clip-level labeling daripada full-video labeling.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.6
Distribusi status yang bersih (tanpa INVALID) menguatkan bahwa kondisi ini sistemik dan stabil, bukan hasil kebetulan beberapa sampel.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.7
Ketiadaan daftar INVALID menegaskan tidak ada item yang perlu di-drop karena mismatch index, sehingga pipeline tidak memerlukan guard tambahan untuk kasus frame out-of-range.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.8
A6 dapat ditutup dengan keputusan “OK untuk subset berlabel”, dan pekerjaan berikutnya bergeser dari “alignment error” menjadi “coverage policy” untuk NO_LABEL dan untuk frame-frame video yang tidak tercakup label.

EDA_STEP_A6_FrameIndex_VideoFra…

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- **Tidak ada masalah INVALID alignment** (0 kasus), jadi tidak ada indikasi label menunjuk frame non-eksis.

  EDA_STEP_A6_FrameIndex_VideoFra…

- **Isu yang tersisa** bukan mismatch indeks, melainkan **coverage**:

1. Ada **39 video NO_LABEL** (tidak punya label sama sekali).

   EDA_STEP_A6_FrameIndex_VideoFra…

2. Dari scatter terlihat label kemungkinan hanya mencakup sebagian awal video (hipotesis kontrak “window labeling”), sehingga frame di luar rentang label **bukan otomatis negatif**.

   EDA_STEP_A6_FrameIndex_VideoFra…

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Bangun dataset training supervised hanya dari video dengan status **VALID** (961 video), dan treat **NO_LABEL (39)** sebagai bucket terpisah (di-drop atau dipakai untuk eksperimen unsupervised/SSL).

  EDA_STEP_A6_FrameIndex_VideoFra…

- Tambahkan aturan saat frame extraction: gunakan **rentang frame berlabel** (mis. 0..max_frameIndex_label) untuk membentuk sample training, agar tidak mencampur frame yang tidak punya annotation. (Motivasi dari pola scatter)

  EDA_STEP_A6_FrameIndex_VideoFra…

Kaitan dengan Modeling

- Training detector/MOT sebaiknya tidak menganggap “absence of label = background” untuk frame di luar coverage label; gunakan sampling hanya pada window berlabel atau definisikan policy negatif secara eksplisit.

  EDA_STEP_A6_FrameIndex_VideoFra…

- Karena alignment aman, Anda bisa lanjut ke analisis coverage severity (A7) dan semantics flag (A8) tanpa risiko frameIndex out-of-range.

▯ STEP A6 — FrameIndex ↔ Video Frame Count Consistency Check

1▯ Insight Detail per Sub-step / Output

▯ Output A6.1 — Video Metadata Load

Fakta Penting

- Metadata video berhasil dimuat untuk **1000 video** (frame_count sebagai ground truth).

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Referensi durasi video sudah lengkap untuk evaluasi konsistensi label-frame.

Ini menunjukkan

- Sisi “video” tidak menjadi bottleneck untuk validasi alignment; semua video punya frame_count.

Resiko

- –

▯ Output A6.2 — Max FrameIndex per Video

Fakta Penting

- Label menghasilkan **1400 video unik** setelah normalisasi `.mov`.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Label memang mencakup video di luar subset train yang tersedia (konsisten dengan A3).

Ini menunjukkan

- Normalisasi key videoName sudah bekerja, dan A6 hanya perlu mengevaluasi subset train (1000 video).

Resiko

- –

Output A6.3 — Label ↔ Video Join

Fakta Penting

- Join menghasilkan **1000 baris** (tepat jumlah video train).

  EDA_STEP_A6_FrameIndex_VideoFra…

- `NaN max_frameIndex_label` dipakai untuk menandai **NO_LABEL**.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Joinability pada level identifier video sudah deterministik dan konsisten untuk evaluasi status per-video.

Ini menunjukkan

- Masalah “semua NO_LABEL” yang terjadi sebelumnya sudah terselesaikan; A6 sekarang mengukur hal yang benar.

Resiko

- –

Output A6.4 — Alignment Status Summary

Fakta Penting

- Status video: **VALID = 961**, **NO_LABEL = 39**, **INVALID = 0**.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Untuk seluruh video yang berlabel, **tidak ada kasus frameIndex label yang melampaui durasi video**.

Ini menunjukkan

- Tidak ada “silent mismatch fatal” tipe: label menunjuk frame yang tidak eksis (untuk subset berlabel).

Resiko

- Untuk 39 video NO_LABEL, pipeline training supervised harus eksplisit memutuskan perlakuannya agar tidak masuk train secara tidak sengaja (bukan risiko alignment, tapi risiko dataset assembly).

□ Output A6.5 — Scatter: FrameIndex vs FrameCount

Fakta Penting

- Plot menunjukkan titik-titik **jauh di bawah garis batas $y \leq x-1$** (tidak ada yang menembus batas).

  EDA_STEP_A6_FrameIndex_VideoFra…

- Secara visual, `max_frameIndex_label` tampak berkisar sekitar **~100–~210**, sementara frame_count video berada di kisaran **~600–~1600**. (terlihat pada page plot)

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Label frameIndex berada jauh di bawah panjang video → label tampaknya hanya mencakup **subset awal frame** dari tiap video, bukan seluruh durasi video.

Ini menunjukkan

- Kontrak label kemungkinan “sequence/window labeling” (mis. label hanya untuk sekitar ~200 frame awal), bukan full-video annotation—namun tetap konsisten secara indeks.

Resiko

- Jika preprocessing/training mengasumsikan “seluruh frame video berlabel”, maka frame-frame setelah `max_frameIndex_label` bisa salah diperlakukan sebagai background/negative secara implisit (label-missing ≠ negative).

□ Output A6.6 — Alignment Status Distribution

Fakta Penting

- Distribusi hanya berisi **VALID dan NO_LABEL**; tidak ada INVALID.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Kondisi “aman dari INVALID” bersifat sistemik, bukan kebetulan sporadis.

Ini menunjukkan

- Secara global, dataset (subset berlabel) memenuhi kontrak alignment yang diperlukan untuk training.

Resiko

- –

□ Output A6.7 — Invalid Alignment Listing

Fakta Penting

- Tercatat eksplisit: **tidak ada video** dengan frameIndex label melebihi decoded frame count.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Tidak diperlukan mekanisme filtering karena INVALID (karena memang 0).

Ini menunjukkan

- Step fail-fast ini lulus: tidak ada kasus label menunjuk frame non-eksis.

Resiko

- –

Output A6.8 — Alignment Decision Summary

Fakta Penting

- Ringkasan keputusan: **VALID 961, INVALID 0, NO_LABEL 39** dan rule keputusan sesuai.

  EDA_STEP_A6_FrameIndex_VideoFra…

Insight

- Keputusan A6 adalah **alignment OK untuk subset berlabel**.

Ini menunjukkan

- Dataset siap lanjut ke langkah berikutnya tanpa perbaikan alignment, cukup penanganan 39 video NO_LABEL sebagai kasus coverage.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step A6.1
Dengan frame_count tersedia untuk 1000 video, A6 punya ground truth yang kuat untuk memastikan label tidak menunjuk ke frame di luar durasi video.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.2
Perhitungan max frameIndex pada label sekaligus menegaskan bahwa label mencakup video yang lebih luas (1400), sehingga evaluasi harus berbasis subset train yang memang ada.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.3
Join yang menghasilkan tepat 1000 baris dan penggunaan NO_LABEL via NaN menunjukkan proses pemetaan video ↔ label sudah stabil dan dapat dipercaya untuk evaluasi status per video.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.4
Hasil status (961 VALID, 39 NO_LABEL, 0 INVALID) menutup risiko fatal “label melampaui durasi video” untuk subset yang berlabel, sambil menandai adanya gap coverage yang harus diperlakukan khusus untuk 39 video.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.5
Scatter memperlihatkan pola penting: walaupun alignment aman, label tampaknya hanya meng-cover sebagian awal video (max_frameIndex sekitar ~200) sementara videonya jauh lebih panjang, sehingga kontrak data lebih mirip window/clip-level labeling daripada full-video labeling.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.6
Distribusi status yang bersih (tanpa INVALID) menguatkan bahwa kondisi ini sistemik dan stabil, bukan hasil kebetulan beberapa sampel.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.7
Ketiadaan daftar INVALID menegaskan tidak ada item yang perlu di-drop karena mismatch index, sehingga pipeline tidak memerlukan guard tambahan untuk kasus frame out-of-range.

EDA_STEP_A6_FrameIndex_VideoFra…

Sub-step A6.8
A6 dapat ditutup dengan keputusan “OK untuk subset berlabel”, dan pekerjaan berikutnya bergeser dari “alignment error” menjadi “coverage policy” untuk NO_LABEL dan untuk frame-frame video yang tidak tercakup label.

EDA_STEP_A6_FrameIndex_VideoFra…

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- **Tidak ada masalah INVALID alignment** (0 kasus), jadi tidak ada indikasi label menunjuk frame non-eksis.

  EDA_STEP_A6_FrameIndex_VideoFra…

- **Isu yang tersisa** bukan mismatch indeks, melainkan **coverage**:

1. Ada **39 video NO_LABEL** (tidak punya label sama sekali).

   EDA_STEP_A6_FrameIndex_VideoFra…

2. Dari scatter terlihat label kemungkinan hanya mencakup sebagian awal video (hipotesis kontrak “window labeling”), sehingga frame di luar rentang label **bukan otomatis negatif**.

   EDA_STEP_A6_FrameIndex_VideoFra…

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Bangun dataset training supervised hanya dari video dengan status **VALID** (961 video), dan treat **NO_LABEL (39)** sebagai bucket terpisah (di-drop atau dipakai untuk eksperimen unsupervised/SSL).

  EDA_STEP_A6_FrameIndex_VideoFra…

- Tambahkan aturan saat frame extraction: gunakan **rentang frame berlabel** (mis. 0..max_frameIndex_label) untuk membentuk sample training, agar tidak mencampur frame yang tidak punya annotation. (Motivasi dari pola scatter)

  EDA_STEP_A6_FrameIndex_VideoFra…

Kaitan dengan Modeling

- Training detector/MOT sebaiknya tidak menganggap “absence of label = background” untuk frame di luar coverage label; gunakan sampling hanya pada window berlabel atau definisikan policy negatif secara eksplisit.

  EDA_STEP_A6_FrameIndex_VideoFra…

- Karena alignment aman, Anda bisa lanjut ke analisis coverage severity (A7) dan semantics flag (A8) tanpa risiko frameIndex out-of-range.

❓ STEP A7 — Video Without Label Analysis (Coverage Severity)

1️⃣ Insight Detail per Sub-step / Output

❓ Output A7.1 — Video Without Label Identification

Fakta Penting

- Total video train: **1000**.
- Video dengan label: **961**.
- Video tanpa label: **39**.

  EDA_STEP_A7_Video_Without_Label…

Insight

- Sekitar **3.9%** dari video train tidak memiliki label sama sekali.

Ini menunjukkan

- Coverage label tidak sempurna, namun gap-nya relatif kecil dan terlokalisasi pada subset tertentu.

Resiko

- Jika video tanpa label tidak difilter secara eksplisit, pipeline supervised training berisiko memasukkan data tanpa anotasi sebagai data negatif secara implisit.

❓ Output A7.2 — Video Metadata with Label Status

Fakta Penting

- Distribusi status metadata: **LABELED = 961**, **NO_LABEL = 39**.

  EDA_STEP_A7_Video_Without_Label…

Insight

- Status label dapat dipetakan secara deterministik ke seluruh metadata video.

Ini menunjukkan

- Analisis perbandingan metadata antar kelompok dapat dilakukan tanpa bias sampling atau kehilangan baris.

Resiko

- –

Output A7.3 — Metadata Statistics (Labeled vs Unlabeled)

Fakta Penting

- **Frame count (mean)**:
  - LABELED: ~**1182** frame
  - NO_LABEL: ~**1107** frame
- **Minimum frame count**:
  - LABELED: **595**
  - NO_LABEL: **313**
- **FPS** kedua kelompok hampir identik (~30 FPS).
- **Resolusi** identik untuk semua video (1280×720).

  EDA_STEP_A7_Video_Without_Label…

Insight

- Video tanpa label **sedikit lebih pendek secara rata-rata**, dan memiliki minimum durasi yang lebih rendah, namun tidak berbeda ekstrem pada FPS maupun resolusi.

Ini menunjukkan

- Ketiadaan label **bukan karena perbedaan format atau kualitas teknis**, melainkan kemungkinan karena kebijakan anotasi atau seleksi subset.

Resiko

- Jika diasumsikan “video tanpa label = data anomali”, analisis bisa keliru karena secara teknis metadata mereka masih sangat mirip dengan video berlabel.

Output A7.4 — Frame Count Distribution (Boxplot)

Fakta Penting

- Distribusi frame count LABELED dan NO_LABEL **saling overlap kuat**.
- NO_LABEL memiliki beberapa video dengan durasi lebih pendek (outlier bawah), namun median mendekati LABELED.

  EDA_STEP_A7_Video_Without_Label…

Insight

- Video tanpa label **bukan kelompok outlier ekstrem**, melainkan berada dalam rentang distribusi yang sama dengan video berlabel.

Ini menunjukkan

- Ketidakhadiran label kemungkinan **tidak berkorelasi kuat dengan durasi video**.

Resiko

- Menghapus video NO_LABEL dengan asumsi “durasi tidak wajar” tidak didukung oleh bukti distribusi.

Output A7.5 — File Size Distribution

Fakta Penting

- Metadata ukuran file (**size_MB**) tidak tersedia, sehingga perbandingan ukuran file tidak dapat dilakukan.

  EDA_STEP_A7_Video_Without_Label…

Insight

- Tidak ada bukti bahwa video tanpa label lebih kecil atau terkompresi secara berbeda.

Ini menunjukkan

- Perbedaan coverage label **tidak dapat dijelaskan dari sisi ukuran file** (berdasarkan data yang tersedia).

Resiko

- Analisis berbasis ukuran file tidak bisa digunakan sebagai dasar keputusan pada dataset ini.

□ Output A7.6 — Video Without Label List

Fakta Penting

- Terdapat **39 nama file video** yang teridentifikasi eksplisit sebagai NO_LABEL.

  EDA_STEP_A7_Video_Without_Label…

Insight

- Subset video tanpa label dapat diisolasi secara eksplisit dan stabil.

Ini menunjukkan

- Perlakuan khusus (drop, split, atau reuse) dapat diterapkan tanpa ambiguitas identitas data.

Resiko

- –

□ Output A7.7 — Coverage Severity Summary

Fakta Penting

- Ringkasan menegaskan kembali: **961 berlabel**, **39 tanpa label**.
- Step ini bersifat **deskriptif saja**, tanpa keputusan final.

  EDA_STEP_A7_Video_Without_Label…

Insight

- Coverage gap terukur dengan jelas dan relatif kecil.

Ini menunjukkan

- Dataset masih layak untuk supervised training, dengan syarat ada kebijakan eksplisit terhadap NO_LABEL.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step A7.1
Identifikasi awal menunjukkan bahwa hanya sebagian kecil video train (39 dari 1000) yang tidak memiliki label, sehingga masalah coverage bersifat terbatas dan terukur, bukan kegagalan sistemik.

EDA_STEP_A7_Video_Without_Label…

Sub-step A7.2
Pemetaan status label ke metadata video berjalan konsisten, memungkinkan analisis perbandingan tanpa kehilangan konteks atau inkonsistensi jumlah data.

EDA_STEP_A7_Video_Without_Label…

Sub-step A7.3
Statistik metadata memperlihatkan bahwa video tanpa label secara teknis sangat mirip dengan video berlabel—FPS dan resolusi identik, dengan perbedaan utama hanya pada durasi rata-rata yang sedikit lebih pendek.

EDA_STEP_A7_Video_Without_Label…

Sub-step A7.4
Visualisasi distribusi durasi menegaskan bahwa video NO_LABEL bukan outlier ekstrem, melainkan masih berada dalam rentang distribusi utama video berlabel, sehingga ketiadaan label tidak berkorelasi kuat dengan durasi.

EDA_STEP_A7_Video_Without_Label…

Sub-step A7.5
Ketiadaan metadata ukuran file membatasi analisis dari sisi storage, namun tidak mengubah kesimpulan utama bahwa coverage gap tidak dapat dijelaskan oleh perbedaan teknis sederhana.

EDA_STEP_A7_Video_Without_Label…

Sub-step A7.6
Daftar eksplisit video tanpa label memberikan dasar yang kuat untuk penanganan terpisah pada tahap preprocessing dan dataset assembly.

EDA_STEP_A7_Video_Without_Label…

Sub-step A7.7
Ringkasan coverage mengkristalkan temuan bahwa masalah utama bukan kualitas video, melainkan kebijakan anotasi, sehingga keputusan selanjutnya bersifat desain pipeline, bukan perbaikan data mentah.

EDA_STEP_A7_Video_Without_Label…

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- **Masalah inti**: tidak semua video train memiliki label (coverage ~96.1%).

  EDA_STEP_A7_Video_Without_Label…

- **Hipotesis paling masuk akal**:
  - Video NO_LABEL dikeluarkan dari proses anotasi karena kebijakan sampling, bukan karena cacat teknis.
- Tidak ada indikasi bahwa video NO_LABEL rusak, format berbeda, atau outlier ekstrem secara metadata.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Video NO_LABEL **harus dikeluarkan dari dataset supervised training** secara eksplisit.
- Alternatif: simpan sebagai split terpisah untuk:
  - pretraining unsupervised / self-supervised,
  - sanity check pipeline video loading,
  - atau eksperimen domain adaptation.

Kaitan dengan Modeling

- Model supervised (deteksi/MOT) sebaiknya hanya dilatih pada **961 video berlabel**.
- Video NO_LABEL tidak boleh dianggap sebagai negative sample default, karena tidak ada bukti bahwa mereka memang “tanpa objek”.

✅ STEP A8 — haveVideo Flag Semantics & Contract Validation

1️⃣ Insight Detail per Sub-step / Output

✅ Output A8.1 — haveVideo Distribution

Fakta Penting

- Distribusi label:
  - `haveVideo = True`: **1,922,517 baris**
  - `haveVideo = False`: **968,329 baris**

    EDA_STEP_A8_haveVideo_Flag_Sema…

Insight

- Flag `haveVideo` digunakan secara masif dan seimbang relatif besar di kedua nilai, sehingga ini bukan flag minor atau noise.

Ini menunjukkan

- `haveVideo` adalah **kontrak semantik penting** di level label, bukan atribut opsional.

Resiko

- Jika flag ini diabaikan, pipeline berisiko mencampur label yang secara eksplisit menyatakan “tidak ada video”.

✅ Output A8.2 — haveVideo × Video Existence Crosstab

Fakta Penting

- Semua baris dengan:
  - `haveVideo = True` → **video_exists_in_train = True**
  - `haveVideo = False` → **video_exists_in_train = False**
- Tidak ada satupun kasus silang (0 mismatch).

  EDA_STEP_A8_haveVideo_Flag_Sema…

Insight

- `haveVideo` **100% konsisten** dengan keberadaan fisik file video di train set.

Ini menunjukkan

- Flag `haveVideo` secara praktis bermakna:
  "apakah file video fisik tersedia", bukan sekadar metadata abstrak.

Resiko

- –

Output A8.3 — haveVideo × Orphan Label Crosstab

Fakta Penting

- Semua orphan label (video tidak ada di train):
  - Terkonsentrasi pada `haveVideo = False`
- Semua label dengan `haveVideo = True` **bukan orphan**.

  EDA_STEP_A8_haveVideo_Flag_Sema…

Insight

- Orphan label **sepenuhnya terjelaskan** oleh flag `haveVideo = False`.

Ini menunjukkan

- Tidak ada "orphan label misterius"; kondisi orphan adalah **state yang disengaja dan ditandai secara eksplisit**.

Resiko

- –

Output A8.4 — haveVideo × Video Without Label Crosstab

Fakta Penting

- Untuk `haveVideo = True`:
  - **961 video** adalah berlabel (video_without_label = False)

- Untuk `haveVideo = False`:
  - **39 video** adalah video train tanpa label (video_without_label = True)
- Tidak ada kombinasi silang yang aneh.

  EDA_STEP_A8_haveVideo_Flag_Sema…

Insight

- Video train tanpa label **selalu** berkorelasi dengan `haveVideo = False`.

Ini menunjukkan

- Gap coverage label (39 video) **bukan error anotasi**, tetapi konsekuensi langsung dari kontrak `haveVideo`.

Resiko

- Jika video tanpa label diperlakukan sebagai “kesalahan data”, itu akan salah kaprah terhadap desain dataset.

□ Output A8.5 — Contract Rule Evaluation

Fakta Penting

- Rule candidates diuji dan seluruhnya tervalidasi oleh crosstab:
  1. `haveVideo == False` & video exists → **tidak terjadi**
  2. `haveVideo == True` & video missing → **tidak terjadi**
  3. `haveVideo == False` & orphan label → **terjadi konsisten**

     EDA_STEP_A8_haveVideo_Flag_Sema…

Insight

- Tidak ditemukan pelanggaran kontrak data terkait `haveVideo`.

Ini menunjukkan

- `haveVideo` dapat digunakan sebagai **aturan filtering yang aman dan deterministik**.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step A8.1
Distribusi besar dan signifikan dari `haveVideo=True/False` menegaskan bahwa flag ini adalah bagian inti dari desain label, bukan atribut sekunder.

EDA_STEP_A8_haveVideo_Flag_Sema…

Sub-step A8.2
Konsistensi sempurna antara `haveVideo` dan keberadaan file video membuktikan bahwa flag ini merepresentasikan kondisi fisik data secara langsung.

EDA_STEP_A8_haveVideo_Flag_Sema…

Sub-step A8.3
Seluruh orphan label berada pada `haveVideo=False`, menegaskan bahwa kondisi orphan adalah state yang disengaja dan sudah diberi penanda eksplisit oleh pembuat dataset.

EDA_STEP_A8_haveVideo_Flag_Sema…

Sub-step A8.4
Korelasi penuh antara video train tanpa label dan `haveVideo=False` mengaitkan temuan A7 secara langsung ke kontrak data, bukan ke masalah kualitas atau kehilangan anotasi.

EDA_STEP_A8_haveVideo_Flag_Sema…

Sub-step A8.5
Validasi rule-based mengukuhkan bahwa `haveVideo` adalah flag kontrak yang konsisten dan dapat dijadikan filter resmi dalam pipeline preprocessing dan dataset assembly.

EDA_STEP_A8_haveVideo_Flag_Sema…

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- **Tidak ditemukan masalah kontrak** pada flag `haveVideo`.

  EDA_STEP_A8_haveVideo_Flag_Sema…

- Video tanpa label (39 video) **bukan kesalahan data**, melainkan konsekuensi eksplisit dari `haveVideo=False`.

- Orphan label sepenuhnya terjelaskan oleh flag ini, sehingga tidak ada risiko silent mismatch pada joinability.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Gunakan aturan keras:
- `hanya gunakan data dengan haveVideo == True`

  untuk membangun dataset supervised.

- Semua label dengan `haveVideo=False` **harus di-drop dari supervised pipeline**, tanpa perlu pengecekan tambahan.

Kaitan dengan Modeling

- Model supervised (deteksi/MOT) **tidak boleh** dilatih pada data `haveVideo=False`.
- Video dengan `haveVideo=False` dapat:
  - diabaikan sepenuhnya, atau
  - dimanfaatkan untuk task non-supervised (pretraining visual, domain shift analysis), **jika dan hanya jika** file videonya tersedia di luar train set ini.

📌 Matriks 1 — Kumpulan Insight Informatif (Disatukan)

Area	Insight Informatif (gabungan lintas step)	Dampak Praktis
Aset dataset (video)	Dataset train berisi **1000 video** `.mov` dan secara file-level **sehat** (tidak ada indikasi corrupt/abnormal).	Pipeline ekstraksi frame & decoding bisa dibuat sederhana (tanpa banyak exception handling).
Sumber label	Label tersedia dalam **CSV & Parquet** dan keduanya dapat dibaca; pemilihan Parquet lebih ke efisiensi eksekusi.	Gunakan Parquet sebagai source utama untuk EDA lanjutan & training pipeline.
Kontrak bbox	Format bbox adalah **pixel absolut** dengan format **xyxy** (bukan normalized).	Preprocessing bbox tidak perlu denormalisasi; perlu clamp ke boundary resolusi.
Kontrak frameIndex	`frameIndex` bersifat **0-based** dan dapat dipakai sebagai indeks frame hasil decode.	Frame extraction & label alignment bisa rule-based (langsung by index).
Kontrak track	`track_id` bersifat **lokal per video** (bukan global).	Saat membuat dataset untuk MOT, identitas track harus dibatasi per video/sequence.
Joinability scope	Label mencakup lebih banyak video daripada subset train; terdapat fenomena label global vs train subset.	Semua operasi training wajib memfilter “subset train yang punya video”.
Coverage train berlabel	Dari 1000 video train: **961 berlabel**, **39 tanpa label**.	Dataset supervised bisa jalan dengan 961 video; 39 video harus diperlakukan khusus.
Alignment label–video	Tidak ada kasus label frame melampaui durasi video (INVALID = 0).	Tidak perlu guard khusus untuk frameIndex out-of-range pada subset berlabel.
Window-level labeling	Scatter menunjukkan label hanya meng-cover **subset awal frame** (sekitar ratusan frame), sementara video jauh lebih panjang.	Training sebaiknya sampling frame hanya pada rentang berlabel; “frame tanpa label” tidak otomatis negatif.
Metadata consistency	Resolusi & FPS konsisten antar video; video NO_LABEL tidak berbeda ekstrem dari video berlabel.	Tidak perlu normalisasi resolusi/FPS; NO_LABEL bukan karena kualitas teknis.
Semantik haveVideo	`haveVideo` **100% konsisten** terhadap keberadaan video fisik dan menjelaskan orphan label & NO_LABEL.	`haveVideo` bisa dijadikan aturan filtering kontrak yang hard.

⚠ Matriks 2 — Kumpulan Masalah (Risk Register Block A)

Observasinya	Resiko (jika tidak ditangani)	Hipotesis
Ada **gap coverage**: 39 video train **tanpa label**.	Jika masuk supervised training, video tanpa anotasi bisa dianggap "background-only" → bias negatif, false negative, dan merusak learning signal.	39 video memang **tidak termasuk subset anotasi** (bukan error file), sesuai kontrak dataset.
Label mencakup video di luar train: terdapat **orphan label** (label untuk video yang tidak ada di train).	Bila tidak difilter, pipeline bisa error saat load video/frame atau membangun sample dari video yang tidak exist.	Label dataset bersifat global, sedangkan folder train adalah subset; orphan label adalah konsekuensi desain split.
Pola label tampak **window-level**: label hanya mencakup sebagian awal video, bukan seluruh durasi.	Jika frame di luar coverage dianggap negatif (karena tidak ada bbox), model akan belajar sinyal salah: "objek hilang tiba-tiba" → menurunkan recall dan merusak asosiasi temporal.	Anotasi dilakukan untuk rentang frame tertentu (mis. clip awal) untuk efisiensi, bukan full-video labeling.
Identitas join bisa rapuh jika `videoName` tidak dinormalisasi (mis. ekstensi `.mov`).	Join mismatch menyebabkan semua video terbaca NO_LABEL (silent failure), sehingga seluruh EDA/training downstream menjadi invalid.	Perbedaan format identifier antara label (tanpa ekstensi) dan file system (dengan ekstensi) perlu normalisasi konsisten.
`track_id` tidak global dan bisa mengalami fragmentasi (track continuity tidak selalu mulus).	Jika diasumsikan track kontinu atau global, training MOT bisa salah menggabungkan identitas antar video/segment → error pada association loss / metric tracking.	ID dibuat per video dan proses anotasi/tracking menghasilkan split/fragment pada beberapa track.
Bergantung pada `haveVideo` untuk kontrak data adalah wajib; mengabaikannya membuat dataset campur aduk antara yang punya video fisik dan yang tidak.	Label tanpa video fisik bisa masuk pipeline dan menyebabkan crash saat frame extraction, atau mengacaukan statistik dataset.	`haveVideo=False` menandai sampel label yang memang tidak memiliki video di split ini; flag ini desain kontrak dataset.