□ STEP B1 — Missingness & Basic Validity Check

EDA_STEP_B1_Missingness_and_Bas…

---

1 □ Insight Detail per Sub-step / Output

□ Output B1.1 — Scope & Column Contract

Fakta Penting

- Total baris label (original): **2,886,916**.
- Baris label joinable (haveVideo=True): **1,919,666**.
- Kolom wajib untuk audit B1 tersedia lengkap: `videoName, frameIndex, id, category, box2d.x1, box2d.y1, box2d.x2, box2d.y2`.
- Missing required columns: **None**.

Insight

- Ruang lingkup audit kualitas label telah ditetapkan secara deterministik pada subset joinable (haveVideo=True) dengan kontrak kolom inti MOT yang lengkap.

Ini menunjukkan

- STEP B1 dapat mengevaluasi missingness dan validitas bbox tanpa bias dari data out-of-scope (haveVideo=False), dan tanpa hambatan struktural berupa kolom inti yang hilang.

Resiko

- –

  EDA_STEP_B1_Missingness_and_Bas…

□ Output B1.2 — Missing Value Summary (Critical Columns)

Fakta Penting

- Missing count untuk seluruh kolom kritikal adalah **0**.
- Missing rate untuk seluruh kolom kritikal adalah **0.0**.

Insight

- Tidak ada missing value pada kolom-kolom inti yang membentuk learning signal untuk detection/tracking.

Ini menunjukkan

- Pipeline preprocessing tidak perlu menambahkan mekanisme imputasi/penanganan missing untuk kolom inti label (setidaknya pada subset joinable haveVideo=True).

Resiko

- –

  EDA_STEP_B1_Missingness_and_Bas…

□ Output B1.3 — Bounding Box Basic Validity

Fakta Penting

- Jumlah bbox dengan `x1 < 0`: **0**.
- Jumlah bbox dengan `y1 < 0`: **0**.
- Jumlah bbox dengan `width <= 0` (x2-x1 <= 0): **0**.
- Jumlah bbox dengan `height <= 0` (y2-y1 <= 0): **0**.
- `any_invalid`: **0**.

Insight

- Tidak terdeteksi bbox invalid secara matematis pada subset joinable.

Ini menunjukkan

- Dataset joinable memiliki integritas geometri bbox yang konsisten terhadap aturan dasar; risiko kegagalan training akibat bbox negatif/nol tidak terindikasi pada tahap ini.

Resiko

- –

  EDA_STEP_B1_Missingness_and_Bas…

□ Output B1.4 — Category / Class Validity

Fakta Penting

- Distribusi kategori (value counts) terobservasi:
  - car **1,473,500**
  - pedestrian **254,450**
  - truck **95,969**
  - bus **38,854**
  - bicycle **19,229**
  - other vehicle **13,271**
  - rider **13,037**
  - motorcycle **7,801**
  - other person **1,648**
  - train **1,053**
  - trailer **854**
- Missing/NaN category rows: **0**.

Insight

- Kolom `category` terisi penuh dan secara operasional membentuk himpunan kelas yang konsisten (tidak ada kelas kosong/NaN).

Ini menunjukkan

- Untuk tahap berikutnya, daftar kelas dapat didefinisikan dari himpunan kategori yang muncul ini, dan tidak ada indikasi label kosong yang bisa memutus training signal.

Resiko

- –

  EDA_STEP_B1_Missingness_and_Bas…

□ Output B1.5 — Sample Problematic Rows

Fakta Penting

- Number of sampled rows: **0**.
- Kolom yang disiapkan untuk audit sample: `name, videoName, frameIndex, id, category, attributes.crowd, attributes.occluded, attributes.truncated, box2d.* , haveVideo`.

Insight

- Tidak ada baris yang terdeteksi memenuhi kriteria “problematic” pada definisi B1 (missing core field / bbox invalid / category missing).

Ini menunjukkan

- Audit manual berbasis sampling tidak diperlukan untuk isu missingness/bbox invalid pada tahap ini, karena tidak ada kandidat baris bermasalah yang terdeteksi oleh rule B1.

Resiko

- –

  EDA_STEP_B1_Missingness_and_Bas…

□ Output B1.6 — Issue Severity Status

Fakta Penting

- Severity status: **minor**.

Insight

- STEP B1 tidak menemukan indikasi masalah dasar yang bersifat blocker untuk masuk tahap berikutnya.

Ini menunjukkan

- Risiko “fatal data quality” pada dimensi missingness dan validitas bbox dasar tidak terindikasi pada subset joinable.

Resiko

- –

  EDA_STEP_B1_Missingness_and_Bas…

2□ Insight Kesimpulan (Naratif — Terhubung)

Sub-step B1.1
Ruang lingkup analisis ditetapkan pada label yang joinable (haveVideo=True) dengan volume 1,919,666 baris dari total 2,886,916, dan seluruh kolom inti yang diperlukan untuk audit kualitas MOT tersedia tanpa kekurangan kolom struktural.

Sub-step B1.2
Pemeriksaan missingness menunjukkan bahwa seluruh kolom kritikal (identifier video, frame index, track id, category, serta koordinat bbox) terisi penuh dengan missing rate 0.0, sehingga tidak ada indikasi gap data pada level field inti.

Sub-step B1.3
Validitas bbox berdasarkan aturan dasar (koordinat negatif serta lebar/tinggi non-positif) tidak menemukan pelanggaran sama sekali, menandakan konsistensi geometri anotasi pada subset joinable.

Sub-step B1.4
Distribusi category terdefinisi dengan baik dan tidak mengandung nilai NaN, sehingga kelas-kelas yang muncul dapat dipakai sebagai himpunan label operasional untuk tahap berikutnya tanpa kebutuhan penanganan "unknown/missing class" pada level ini.

Sub-step B1.5
Karena tidak ditemukan kandidat baris bermasalah, sampling baris untuk audit manual tidak menghasilkan contoh, yang konsisten dengan hasil missingness dan bbox-validity yang bersih.

Sub-step B1.6
Severity ditetapkan sebagai minor, sehingga STEP B1 secara keseluruhan tidak memberikan sinyal adanya blocker kualitas label dasar untuk melanjutkan EDA ke tahap berikutnya.

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Tidak terdeteksi masalah missingness pada kolom inti.
- Tidak terdeteksi bbox invalid (negatif atau width/height non-positif).
- Tidak terdeteksi category NaN/missing.
- Dengan demikian, tidak ada risiko/hipotesis masalah pada ruang lingkup STEP B1.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Tidak diperlukan imputasi untuk kolom inti label (`videoName`, `frameIndex`, `id`, `category`, `box2d.*`) pada subset joinable (haveVideo=True).
- Tidak diperlukan rule cleaning untuk kasus bbox negatif atau width/height non-positif pada tahap ini.

- Daftar kelas dapat dibentuk dari himpunan category yang terobservasi (11 kategori) sebagai baseline mapping label→id.

Kaitan dengan Modeling

- Training detection/tracking dapat berasumsi bahwa label joinable memiliki koordinat bbox yang valid secara matematis dan kategori terisi penuh, sehingga error awal training akibat label kosong/invalid bbox tidak terindikasi.
- Distribusi kategori menunjukkan dominasi kelas `car`, sehingga isu imbalance bersifat kuantitatif (long-tail) dan relevan untuk dievaluasi di step distribusi berikutnya, bukan sebagai masalah validitas dasar pada STEP B1.

□ STEP B2 — Duplicate Annotation & Key Consistency

EDA_STEP_B2_Duplicate_Annotatio…

1□ Insight Detail per Sub-step / Output

□ Output B2.1 — Scope & Logical Key Definition

Fakta Penting

- Jumlah baris label joinable (haveVideo=True): **1,919,666**.
- Key logis yang digunakan untuk audit konsistensi: **(videoName, frameIndex, id)**.
- `id` (track_id) telah tervalidasi bersifat **lokal per video** (hasil Block A).

Insight

- Ruang lingkup dan definisi key logis untuk audit duplikasi telah ditetapkan secara eksplisit dan konsisten dengan kontrak MOT.

Ini menunjukkan

- Setiap evaluasi duplikasi dan konflik kunci dapat dilakukan secara deterministik tanpa ambiguitas lintas video atau lintas track.

Resiko

- –

  EDA_STEP_B2_Duplicate_Annotatio…

□ Output B2.2 — Literal Duplicate Row Check

Fakta Penting

- Jumlah literal duplicate rows (seluruh kolom identik): **0**.
- Duplicate rate: **0.000000**.

Insight

- Tidak ditemukan duplikasi baris identik secara literal pada subset label joinable.

Ini menunjukkan

- Dataset tidak mengandung copy baris mentah (byte-level duplication) yang dapat menyebabkan model “belajar dua kali” dari anotasi yang persis sama.

Resiko

- –

  EDA_STEP_B2_Duplicate_Annotatio…

□ Output B2.3 — Logical Key Uniqueness Check

Fakta Penting

- Jumlah pelanggaran key logis (videoName, frameIndex, id): **0**.
- Tidak ada grup key dengan ukuran > 1.

Insight

- Kombinasi (video, frame, track_id) bersifat unik untuk seluruh label joinable.

Ini menunjukkan

- Tidak ada konflik logis di mana satu track direpresentasikan lebih dari satu kali pada frame yang sama dengan key identik.

Resiko

- –

  EDA_STEP_B2_Duplicate_Annotatio…

□ Output B2.4 — Multiple BBox per Track per Frame

Fakta Penting

- Jumlah (video, frame, track_id) dengan lebih dari satu bbox: **0**.

Insight

- Setiap track memiliki tepat satu bounding box per frame.

Ini menunjukkan

- Anotasi tracking mengikuti kontrak MOT standar dan tidak mengandung konflik internal berupa multi-bbox untuk satu track dalam satu frame.

Resiko

- –

  EDA_STEP_B2_Duplicate_Annotatio…

□ Output B2.5 — Conflict Key Samples

Fakta Penting

- Jumlah baris konflik yang ter-sampling: **0**.
- Tidak ada kolom konflik yang perlu ditampilkan.

Insight

- Tidak ada kandidat konflik kunci yang memerlukan audit manual.

Ini menunjukkan

- Seluruh hasil pada sub-step sebelumnya konsisten dan tidak menyisakan kasus abu-abu untuk inspeksi tambahan.

Resiko

- –

  EDA_STEP_B2_Duplicate_Annotatio…

□ Output B2.6 — Issue Severity & Treatment Hint

Fakta Penting

- Severity status: **minor**.
- Treatment hint: **ignore**.

Insight

- STEP B2 tidak mengindikasikan adanya masalah duplikasi atau konflik kunci yang bersifat signifikan.

Ini menunjukkan

- Dari perspektif duplikasi dan konsistensi key, dataset joinable siap digunakan untuk tahap berikutnya tanpa kebutuhan intervensi cleaning khusus.

Resiko

- –

  EDA_STEP_B2_Duplicate_Annotatio…

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step B2.1
Penetapan scope pada label joinable serta definisi key logis (videoName, frameIndex, id) memberikan dasar yang jelas dan konsisten untuk seluruh audit duplikasi pada STEP B2.

Sub-step B2.2
Pemeriksaan duplikasi literal memastikan bahwa dataset tidak mengandung baris identik secara byte-level, sehingga risiko pembelajaran redundan akibat copy data mentah tidak terindikasi.

Sub-step B2.3
Audit keunikan key logis menunjukkan bahwa setiap track pada setiap frame direpresentasikan secara unik, menutup kemungkinan konflik logis akibat duplikasi key.

Sub-step B2.4
Validasi satu bbox per track per frame menegaskan bahwa anotasi tracking mengikuti kontrak MOT yang diharapkan dan tidak mengandung representasi ganda untuk satu objek.

Sub-step B2.5
Ketiadaan sample konflik mengonfirmasi bahwa seluruh hasil audit konsisten dan tidak memerlukan inspeksi manual lanjutan.

Sub-step B2.6
Severity ditetapkan sebagai minor, mengindikasikan bahwa tidak ada blocker pada dimensi duplikasi dan konsistensi kunci untuk melanjutkan EDA ke tahap berikutnya.

3️⃣ Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Tidak terdeteksi duplikasi baris literal.
- Tidak terdeteksi pelanggaran keunikan key logis (video, frame, track_id).
- Tidak terdeteksi multi-bbox pada satu track dalam satu frame.
- Dengan demikian, tidak ada risiko atau hipotesis masalah pada ruang lingkup STEP B2.

4️⃣ Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Tidak diperlukan proses deduplikasi baris atau penanganan konflik key pada tahap preprocessing.
- Key logis (videoName, frameIndex, id) dapat digunakan secara langsung sebagai constraint integritas data pada pipeline selanjutnya.

Kaitan dengan Modeling

- Model detection dan tracking tidak berisiko menerima sinyal ganda dari anotasi duplikat pada frame yang sama.
- Evaluasi tracking dapat mengasumsikan satu representasi bbox per track per frame, sehingga metrik berbasis track (IDF1, MOTA, dsb.) tidak terdistorsi oleh konflik anotasi.

□ STEP B3 — Bounding Box Geometry Sanity

EDA_STEP_B3_Bounding_Box_Geomet…

---

1 □ Insight Detail per Sub-step / Output

□ Output B3.1 — Geometry Feature Construction

Fakta Penting

- Jumlah bbox yang dievaluasi (haveVideo=True): **1,919,666**.
- Fitur geometri yang diturunkan:
  - `bbox_w = x2 - x1`
  - `bbox_h = y2 - y1`
  - `bbox_area = bbox_w × bbox_h`
  - `aspect_ratio = bbox_w / bbox_h`
- Tidak ada mutasi terhadap label asli (EDA-only).

Insight

- Seluruh bbox pada subset joinable telah diproyeksikan ke fitur geometri dasar yang diperlukan untuk audit kewajaran fisik.

Ini menunjukkan

- STEP B3 dapat mengevaluasi kewajaran ukuran dan bentuk bbox secara menyeluruh tanpa ketergantungan pada asumsi implisit atau preprocessing awal.

Resiko

- –

  EDA_STEP_B3_Bounding_Box_Geomet…

□ Output B3.2 — Bounding Box Area Distribution (Stats)

Fakta Penting

- Statistik luas bbox (pixels²):
  - min: **27.12**
  - 1%: **163.71**
  - 5%: **296.84**

    - median (50%): **2,312.38**
    - 90%: **30,296.31**
    - 95%: **63,374.07**
    - 99%: **164,060.81**
    - max: **769,838.30**
- Rata-rata area: **12,485.52**, dengan standar deviasi **32,613.26**.

Insight

- Distribusi area bbox sangat skewed dengan rentang luas yang lebar, dari objek sangat kecil hingga objek sangat besar.

Ini menunjukkan

- Dataset mengandung proporsi signifikan objek kecil (area < ~300 px² pada 5% terbawah) serta ekor panjang objek besar, yang berpotensi memengaruhi sensitivitas model terhadap skala objek.

Resiko

- –

  EDA_STEP_B3_Bounding_Box_Geomet…

□ Output B3.3 — Bounding Box Area Distribution (Histogram, Log Scale)

Fakta Penting

- Histogram log-scale menunjukkan distribusi **long-tail** pada area bbox.
- Kepadatan tinggi berada pada area kecil, dengan frekuensi menurun tajam seiring bertambahnya luas bbox.

Insight

- Objek berukuran kecil mendominasi jumlah anotasi, sementara objek besar relatif jarang.

Ini menunjukkan

- Potensi tantangan pada deteksi small object, terutama jika resolusi input atau strategi augmentasi tidak memadai.

Resiko

- –

EDA_STEP_B3_Bounding_Box_Geomet…

□ Output B3.4 — Aspect Ratio Distribution (Stats)

Fakta Penting

- Statistik aspect ratio (w/h):
  - min: **0.033**
  - 1%: **0.275**
  - median (50%): **1.200**
  - 90%: **2.159**
  - 95%: **2.696**
  - 99%: **4.143**
  - max: **29.57**
- Mean aspect ratio: **1.307**.

Insight

- Mayoritas bbox memiliki aspect ratio moderat (sekitar 1–2), dengan sebagian kecil bbox yang sangat ramping atau sangat melebar.

Ini menunjukkan

- Distribusi bentuk bbox secara umum realistis, namun terdapat outlier bentuk ekstrem yang kemungkinan berkaitan dengan kelas tertentu atau kondisi visual khusus.

Resiko

- –

EDA_STEP_B3_Bounding_Box_Geomet…

□ Output B3.5 — Bounding Box Aspect Ratio Distribution (Histogram)

Fakta Penting

- Histogram menunjukkan konsentrasi besar pada aspect ratio rendah–menengah.
- Ekor panjang hingga >10 menunjukkan keberadaan bbox dengan bentuk ekstrem.

Insight

- Sebagian kecil anotasi memiliki bentuk yang sangat tidak seimbang (sangat tinggi/kurus atau sangat lebar).

Ini menunjukkan

- Perlu kewaspadaan terhadap potensi shape bias atau kesulitan prediksi pada bbox dengan bentuk ekstrem, terutama untuk model dengan asumsi anchor tertentu.

Resiko

- –

  EDA_STEP_B3_Bounding_Box_Geomet…

□ Output B3.6 — Extreme Geometry Flags (EDA-Only)

Fakta Penting

- Threshold berbasis kuantil:
  - Area kecil (≤1%): **≤163.71 px²**
  - Area besar (≥99%): **≥164,060.81 px²**
  - Aspect ratio ekstrem: **≤0.275** atau **≥4.143**
- Persentase bbox ter-flag:
  - small_area: **~1.00%**
  - large_area: **~1.00%**
  - extreme_aspect_ratio: **~2.00%**

Insight

- Proporsi bbox ekstrem secara statistik relatif kecil namun tidak dapat diabaikan.

Ini menunjukkan

- Dataset memiliki ekor ekstrem yang terdefinisi jelas; keberadaan bbox sangat kecil dan bentuk ekstrem perlu dipertimbangkan dalam desain preprocessing dan konfigurasi model.

Resiko

- –

  EDA_STEP_B3_Bounding_Box_Geomet…

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step B3.1
Seluruh bbox pada subset joinable berhasil dipetakan ke fitur geometri dasar, menyediakan fondasi kuantitatif untuk mengevaluasi kewajaran ukuran dan bentuk anotasi.

Sub-step B3.2
Distribusi luas bbox memperlihatkan rentang yang sangat lebar dengan dominasi objek kecil, mengindikasikan bahwa skala objek merupakan karakteristik penting dataset ini.

Sub-step B3.3
Visualisasi log-scale menegaskan pola long-tail pada area bbox, di mana sebagian besar anotasi berada pada area kecil dan hanya sedikit bbox yang sangat besar.

Sub-step B3.4
Distribusi aspect ratio menunjukkan bahwa mayoritas bbox memiliki bentuk realistis, namun terdapat outlier dengan rasio ekstrem yang berpotensi menjadi sumber kesulitan prediksi.

Sub-step B3.5
Histogram aspect ratio memperjelas keberadaan ekor panjang, mengonfirmasi bahwa meskipun jarang, bbox dengan bentuk sangat ekstrem memang ada di dataset.

Sub-step B3.6
Flagging berbasis kuantil menandai sekitar 1–2% bbox sebagai ekstrem dari sisi ukuran atau bentuk, memberikan batas indikatif untuk analisis lanjutan tanpa melakukan cleaning pada tahap ini.

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Dataset mengandung proporsi signifikan objek kecil (area sangat rendah).
- Terdapat bbox dengan bentuk ekstrem (aspect ratio sangat kecil atau sangat besar).
- Kondisi ini bersifat distribusional, bukan kesalahan anotasi eksplisit, dan berpotensi memengaruhi performa deteksi terutama pada skala kecil.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Threshold kuantil (1%–99%) dapat dijadikan referensi awal untuk analisis lanjutan (mis. evaluasi dampak clipping atau filtering), tanpa langsung diterapkan sebagai aturan cleaning.
- Informasi dominasi objek kecil relevan untuk keputusan resolusi input dan strategi augmentasi berbasis skala.

Kaitan dengan Modeling

- Model perlu sensitif terhadap small object, mengingat median area bbox relatif kecil.
- Distribusi aspect ratio yang lebar mengindikasikan perlunya arsitektur atau konfigurasi yang robust terhadap variasi bentuk objek, serta evaluasi khusus pada bbox ekstrem sebagai bagian dari analisis error di tahap modeling.

□ STEP B4 — Track Continuity & Temporal Consistency

EDA_STEP_B4_Track_Continuity_an…

1□ Insight Detail per Sub-step / Output

□ Output B4.1 — Scope & Track Key Contract

Fakta Penting

- Jumlah baris label yang dianalisis (haveVideo=True): **1,919,666**.
- Key tracking didefinisikan sebagai **(videoName, id)**.
- Seluruh kolom yang dibutuhkan (`videoName`, `frameIndex`, `id`, `category`) tersedia.
- Analisis temporal dibatasi hanya pada **frame yang berlabel**.

Insight

- Ruang lingkup dan kontrak key tracking telah ditetapkan dengan jelas dan konsisten dengan hasil Block A.

Ini menunjukkan

- Evaluasi kontinuitas temporal dapat dilakukan secara valid tanpa risiko konflik lintas video atau track_id global.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

□ Output B4.2 — Track Length per (video, track_id)

Fakta Penting

- Jumlah track total: **75,885**.
- Statistik panjang track (dalam frame):
  - min: **1**
  - 1%: **1**
  - 5%: **2**
  - 10%: **3**

- median (50%): **15**
- mean: **25.30**
- 90%: **61**
- 95%: **92**
- 99%: **165**
- max: **263**

Insight

- Distribusi panjang track sangat bervariasi, dengan median relatif pendek dibandingkan ekor panjang pada track yang sangat panjang.

Ini menunjukkan

- Dataset mengandung kombinasi track yang sangat pendek dan track yang sangat panjang, yang berpotensi mencerminkan variasi tingkat kesulitan tracking antar objek dan antar video.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

□ Output B4.3 — Track Length Distribution (Histogram, Log Scale)

Fakta Penting

- Histogram log-scale (halaman 3) menunjukkan:
  - Kepadatan tinggi pada track pendek.
  - Penurunan frekuensi secara gradual seiring bertambahnya panjang track.
  - Ekor panjang hingga >200 frame.

Insight

- Track pendek secara numerik mendominasi jumlah track, meskipun sebagian kecil track memiliki durasi sangat panjang.

Ini menunjukkan

- Secara agregat, tracking cenderung terfragmentasi pada sebagian objek, sementara objek tertentu berhasil dilacak dalam durasi panjang secara stabil.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

□ Output B4.4 — Short Track Prevalence (1–2 frames)

Fakta Penting

- Persentase track:
  - Panjang = 1 frame: **2.59%**
  - Panjang = 2 frame: **4.11%**
  - Panjang ≤ 2 frame: **6.70%**
  - Panjang ≤ 5 frame: **19.66%**

Insight

- Sekitar **6–7%** track sangat pendek (≤2 frame), dan hampir **20%** track berdurasi sangat singkat (≤5 frame).

Ini menunjukkan

- Terdapat tingkat fragmentasi tracking yang moderat, namun tidak dominan secara ekstrem pada subset joinable.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

□ Output B4.5 — Frame Gap Analysis (Within Track)

Fakta Penting

- Total selisih frame (`diff`): **1,843,781**.
- Gap > 1 frame:
  - Jumlah: **26,446**
  - Persentase: **1.43%**
- Statistik gap:
  - median diff: **1**
  - 99% diff: **3**
  - max gap: **169**

- Untuk gap > 1:
  - median: **4**
  - 90%: **12**
  - 99%: **52**

Insight

- Mayoritas track memiliki frameIndex berurutan, dengan gap jarang dan sebagian besar kecil.

Ini menunjukkan

- Diskontinuitas temporal memang ada, tetapi bersifat minor dan tidak mendominasi keseluruhan track; gap ekstrem hanya terjadi pada sebagian kecil kasus.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

□ Output B4.6 — Extreme Track Listing (Too Short / Too Long)

Fakta Penting

- Track terlalu pendek (≤2 frame):
  - Jumlah: **5,086**
  - Persentase: **6.70%**
- Track terlalu panjang (≥ p99.5 ≈ 198 frame):
  - Jumlah: **388**
  - Persentase: **0.51%**
- Contoh track ekstrem ditampilkan pada halaman 6.

Insight

- Track ekstrem berada pada proporsi kecil dari keseluruhan populasi track.

Ini menunjukkan

- Sebagian besar track berada dalam rentang durasi yang wajar, dengan outlier pendek dan panjang yang terdefinisi jelas.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

□ Output B4.7 — Temporal Consistency Severity Status

Fakta Penting

- Severity status: **minor**.
- Heuristik:
  - major jika track_len ≤ 2 ≥ 20% **atau** gap>1 ≥ 10%.

Insight

- Indikator fragmentasi dan diskontinuitas temporal berada jauh di bawah ambang “major”.

Ini menunjukkan

- Dari perspektif kontinuitas temporal, kualitas tracking label masih berada dalam batas yang dapat diterima untuk melanjutkan tahap berikutnya.

Resiko

- –

  EDA_STEP_B4_Track_Continuity_an…

2□ Insight Kesimpulan (Naratif — Terhubung)

Sub-step B4.1
Ruang lingkup dan key tracking ditetapkan secara eksplisit pada label joinable, memastikan bahwa seluruh analisis temporal berjalan pada domain yang valid dan konsisten.

Sub-step B4.2
Distribusi panjang track menunjukkan median relatif pendek dengan ekor panjang, menandakan variasi besar dalam durasi tracking antar objek.

Sub-step B4.3
Visualisasi histogram menegaskan dominasi track pendek secara numerik, namun tetap mempertahankan keberadaan track panjang yang stabil.

Sub-step B4.4
Prevalensi track sangat pendek berada di kisaran moderat dan tidak mencapai tingkat yang mengindikasikan fragmentasi ekstrem.

Sub-step B4.5
Analisis gap frame menunjukkan bahwa mayoritas track memiliki kontinuitas frame yang baik, dengan gap jarang dan umumnya kecil.

Sub-step B4.6
Track ekstrem (sangat pendek atau sangat panjang) teridentifikasi dengan jelas namun hanya mencakup sebagian kecil populasi track.

Sub-step B4.7
Severity ditetapkan sebagai minor, mengindikasikan bahwa secara keseluruhan kontinuitas temporal anotasi tracking berada pada tingkat yang dapat diterima.

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Terdapat track berdurasi sangat pendek (≤2 frame) dalam proporsi terbatas.
- Terdapat diskontinuitas frame (gap >1) pada sebagian kecil track.
- Pola ini lebih bersifat karakteristik distribusi tracking daripada indikasi kesalahan anotasi sistemik.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Track sangat pendek (≤2 frame) dapat dipertimbangkan sebagai kandidat analisis lanjutan (mis. untuk evaluasi dampak filtering), tanpa diterapkan sebagai aturan drop pada tahap ini.
- Informasi gap frame dapat digunakan sebagai referensi untuk strategi sampling temporal atau windowing pada tahap selanjutnya.

Kaitan dengan Modeling

- Model tracking perlu cukup robust terhadap fragmentasi ringan dan gap kecil yang muncul secara alami pada data.
- Evaluasi berbasis track (mis. IDF1, MOTA) sebaiknya mempertimbangkan keberadaan track pendek sebagai bagian dari karakteristik dataset, bukan sebagai noise murni.

□ STEP B5 — Label Noise & Ambiguity (Early Signal)

EDA_STEP_B5_Label_Noise_and_Amb…

1□ Insight Detail per Sub-step / Output

□ Output B5.1 — Scope & Noise Signal Definition

Fakta Penting

- Jumlah baris label yang dianalisis (haveVideo=True): **1,919,666**.
- Key tracking: **(videoName, id)**.
- Sinyal noise yang dieksplorasi:
  - **Multiple classes within a single track**
  - **High IoU overlap between different classes**
- Kolom wajib tersedia lengkap (Missing required columns: **None**).

Insight

- Ruang lingkup dan definisi sinyal noise telah ditetapkan dengan jelas pada domain joinable sehingga indikasi noise yang diukur tidak terkontaminasi oleh baris out-of-scope.

Ini menunjukkan

- STEP B5 dapat bertindak sebagai “early warning system” yang konsisten dengan kontrak tracking: track-id lokal per video, dan evaluasi dilakukan pada subset label yang memang dapat dilatih.

Resiko

- –

  EDA_STEP_B5_Label_Noise_and_Amb…

□ Output B5.2 — Multiple Class per Track

Fakta Penting

- Total tracks: **75,885**.
- Tracks dengan >1 class: **0**.

- Rate: **0.0000%**.

Insight

- Tidak ditemukan track yang berganti class sepanjang durasi track pada dataset joinable.

Ini menunjukkan

- Label kategori bersifat stabil per track, sehingga indikasi noise berupa “class switching dalam track” tidak terdeteksi pada tahap ini.

Resiko

- –

  EDA_STEP_B5_Label_Noise_and_Amb…

□ Output B5.3 — Class Change Frequency

Fakta Penting

- Statistik `num_class_changes` per track:
  - count: **75,885**
  - mean: **1.0**
  - std: **0.0**
  - min=median=max: **1.0**

Insight

- Distribusi `num_class_changes` bersifat degeneratif (konstan).

Ini menunjukkan

- Implementasi metrik “class change count” pada output ini secara operasional menghasilkan nilai konstan untuk seluruh track; sehingga metrik ini, dalam bentuk saat ini, tidak membedakan track stabil vs track berubah kelas.

Resiko

- **–**

  EDA_STEP_B5_Label_Noise_and_Amb…

Catatan teknis (tetap netral): karena B5.2 mendeteksi 0 multi-class track, maka secara konsep “jumlah pergantian class” seharusnya 0 untuk semua track. Output konstan di 1.0 mengindikasikan definisi `class_change` menghitung peristiwa pertama sebagai perubahan (efek `prev_category` NaN) atau aturan agregasi yang membuat baseline 1. Ini tidak mempengaruhi kesimpulan B5.2, tetapi membuat metrik B5.3 tidak bisa dipakai sebagai sinyal noise dalam bentuk sekarang.

□ Output B5.4 — High IoU Overlap Across Classes (Sampled)

Fakta Penting

- Sampling: **2%** random dari joinable labels.
- Jumlah pasangan bbox beda kelas dengan IoU ≥ 0.7 terdeteksi: **0**.
- Tidak ada distribusi IoU yang dicetak karena tidak ada kandidat.

Insight

- Pada sampel ini, tidak ada indikasi overlap ekstrem antar class berbeda (yang biasa menjadi sinyal ambiguity atau label conflict).

Ini menunjukkan

- Indikasi “bbox overlap tinggi beda class” tidak terdeteksi pada evaluasi sampled; jika ada, kemungkinan bersifat jarang atau berada di bawah threshold IoU yang digunakan.

Resiko

- –

  EDA_STEP_B5_Label_Noise_and_Amb…

□ Output B5.5 — Class Ambiguity Ranking

Fakta Penting

- Output class ambiguity counts adalah series kosong (tidak ada multi-class track).

Insight

- Tidak ada basis data untuk melakukan ranking class ambiguity berbasis multi-class track.

Ini menunjukkan

- Dengan definisi ambiguity saat ini (multi-class per track), dataset tidak menunjukkan class yang sering ambigu, karena kasus ambiguity tersebut tidak muncul.

Resiko

- –

  EDA_STEP_B5_Label_Noise_and_Amb…

□ Output B5.6 — Noise Indicator Summary & Severity

Fakta Penting

- `multi_class_track_rate_percent`: **0.0**
- `high_iou_overlap_count_sample`: **0**
- `severity_status`: **minor**

Insight

- Sinyal noise/ambiguity yang ditargetkan pada STEP B5 tidak terindikasi pada domain joinable (dengan sampling untuk IoU).

Ini menunjukkan

- Dari dua indikator awal yang diuji (class switching dalam track & high IoU beda kelas), dataset terlihat stabil pada level definisi yang digunakan.

Resiko

- –

  EDA_STEP_B5_Label_Noise_and_Amb…

2□ Insight Kesimpulan (Naratif — Terhubung)

Sub-step B5.1
Analisis noise dilakukan pada label joinable dengan key tracking (videoName, id), sehingga sinyal yang dievaluasi berada pada domain yang relevan untuk training tracking.

Sub-step B5.2
Pemeriksaan jumlah kelas unik per track tidak menemukan track multi-class, sehingga indikasi "track berubah kelas di tengah jalan" tidak terdeteksi.

Sub-step B5.3
Metrik jumlah perubahan kelas per track menghasilkan nilai konstan (1.0) untuk seluruh track, sehingga pada bentuk saat ini metrik tersebut tidak berfungsi sebagai sinyal pembeda untuk noise; namun hal ini tetap konsisten dengan hasil B5.2 yang tidak menemukan multi-class track.

Sub-step B5.4
Pemeriksaan overlap IoU tinggi antar class berbeda pada sampel 2% tidak menemukan kandidat, sehingga indikasi ambiguity berupa konflik overlap beda kelas tidak terlihat pada evaluasi sampled.

Sub-step B5.5
Karena tidak ada track multi-class, ranking class ambiguity berbasis multi-class track tidak dapat dibentuk dan output menjadi kosong.

Sub-step B5.6
Ringkasan indikator menetapkan severity minor, menandakan tidak ada sinyal awal noise/ambiguity dari definisi yang digunakan pada STEP B5.

---

3 Penjelasan Masalah yang Terjadi (Rangkuman Risiko & Hipotesis)

- Tidak ada indikasi noise dari:
  - track yang berganti class
  - overlap bbox beda class dengan IoU ≥ 0.7 (pada sampel 2%)
- Terdapat satu isu teknis pada metrik B5.3: output `num_class_changes` konstan 1.0 untuk semua track, sehingga metrik tersebut tidak informatif dalam bentuk sekarang.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Tidak ada kebutuhan preprocessing khusus untuk menangani:
  - track yang berubah kelas
  - konflik overlap beda class (berdasarkan indikator dan threshold yang diuji)

- Jika metrik “class change count” ingin dipakai pada tahap lanjutan, definisinya perlu diselaraskan agar baseline track stabil menghasilkan 0 (bukan 1), namun ini bersifat perbaikan sinyal EDA, bukan kebutuhan cleaning data.

Kaitan dengan Modeling

- Dengan tidak ditemukannya class-switch per track, training tracker dapat mengasumsikan **kategori stabil per track** (sesuai kontrak label).
- Tidak adanya high IoU beda class pada sampel mendukung asumsi bahwa konflik label antar class akibat overlap ekstrem tidak dominan pada dataset ini (dengan definisi threshold yang digunakan), sehingga fokus modeling bisa dialihkan ke tantangan lain yang sudah terindikasi pada Step sebelumnya (mis. distribusi skala bbox dan fragmentasi track).

□ STEP B6 — Frame Boundary & Clipping Validity (BBox Out-of-Frame)

EDA_STEP_B6_Frame_Boundary_and_…

1□ Insight Detail per Sub-step / Output

□ Output B6.1 — Scope & Metadata Discovery

Fakta Penting

- Jumlah baris label dianalisis (haveVideo=True): **1,919,666**.
- Seluruh kolom label bbox tersedia lengkap.
- **Metadata resolusi frame (W/H) tidak ditemukan sama sekali**:
  - `Metadata source detected`: **None**
  - `Metadata rows (unique videoName)`: **0**

Insight

- Tidak tersedia sumber metadata resolusi video yang dapat dipakai untuk validasi batas frame.

Ini menunjukkan

- Pada kondisi saat ini, **validasi bbox terhadap batas frame tidak dapat dilakukan**, terlepas dari kualitas anotasi bbox itu sendiri.

Resiko

- –

□ Output B6.2 — Metadata Join Coverage

Fakta Penting

- Total label rows: **1,919,666**
- Rows dengan metadata W/H: **0**
- Coverage metadata: **0.0%**

Insight

- Tidak satu pun bbox joinable dapat dipetakan ke resolusi frame.

Ini menunjukkan

- Seluruh analisis boundary validity berbasis W/H menjadi **non-operational**.

Resiko

- –

□ Output B6.3 — BBox Boundary Issue Summary (Rule-Based)

Fakta Penting

- Semua rule boundary (`x1<0`, `y1<0`, `x2>W`, `y2>H`) menghasilkan:
  - count = **0**
  - rate = **0.0**

Insight

- Nilai nol **bukan berarti bbox valid**, tetapi karena **tidak ada baris yang memenuhi syarat evaluasi (W/H missing)**.

Ini menunjukkan

- Output boundary check **tidak informatif** dalam kondisi metadata kosong.

Resiko

- –

□ Output B6.4 — Overshoot Severity Distribution

Fakta Penting

- Tidak ada statistik overshoot yang dapat dihitung.
- Seluruh deskripsi dan quantile bertuliskan *"No rows with metadata (W/H)"*.

Insight

- Tidak ada dasar kuantitatif untuk menilai seberapa jauh bbox melampaui batas frame.

Ini menunjukkan

- Analisis severity overshoot **sepenuhnya terblokir** oleh absennya metadata resolusi.

Resiko

- –

□ Output B6.5 — Severity Buckets & Treatment Hint

Fakta Penting

- Semua bucket (minor / moderate / severe / in_frame) = **0**.
- Treatment hint:
  - clip_candidates = **0**
  - drop_candidates = **0**

Insight

- Bucket kosong merupakan konsekuensi langsung dari **tidak adanya evaluasi boundary**.

Ini menunjukkan

- Tidak mungkin menyimpulkan apakah bbox perlu di-clip atau di-drop tanpa resolusi frame.

Resiko

- –

□ Output B6.6 — Breakdown by Category & Video

Fakta Penting

- Tidak ada breakdown per category maupun per video.
- Seluruh tabel menyatakan *"No rows with metadata (W/H)"*.

Insight

- Tidak bisa diidentifikasi class atau video yang berpotensi bermasalah secara boundary.

Ini menunjukkan

- Risiko boundary bersifat **unknown**, bukan **absent**.

Resiko

- –

□ Output B6.7 — Sample Out-of-Frame Rows

Fakta Penting

- Sample rows count: **0**

Insight

- Tidak tersedia contoh bbox out-of-frame untuk audit manual.

Ini menunjukkan

- Ketidakmampuan audit bukan karena data bersih, melainkan karena **kontrak informasi tidak lengkap**.

Resiko

- –

□ Output B6.8 — Boundary Validity Severity Status

Fakta Penting

- **Severity status: FATAL**
- Heuristik:
  - fatal jika **tidak ada metadata W/H sama sekali**

Insight

- Status fatal dipicu **bukan oleh kesalahan anotasi bbox**, tetapi oleh **ketiadaan metadata resolusi frame**.

Ini menunjukkan

- Dari perspektif *Label Health & Data Quality*, dataset **belum memenuhi kontrak minimum** untuk validasi fisik bbox.

Resiko

- –

---

2 Insight Kesimpulan (Naratif — Terhubung)

Sub-step B6.1–B6.2
Tidak adanya metadata resolusi video menyebabkan seluruh label joinable kehilangan konteks spasial absolut terhadap frame.

Sub-step B6.3–B6.5
Rule boundary check, overshoot severity, dan bucket treatment tidak dapat dievaluasi karena prasyarat W/H tidak terpenuhi.

Sub-step B6.6–B6.7
Tanpa metadata, tidak mungkin melakukan analisis diferensial per class atau per video, maupun audit contoh kasus.

Sub-step B6.8
Severity ditetapkan sebagai **fatal**, menandakan kegagalan kontrak data yang mendasar untuk pipeline detection/tracking.

---

3 Penjelasan Masalah yang Terjadi (Observasi, Risiko, Hipotesis)

Observasi

- Metadata resolusi frame (width/height) **tidak tersedia** di lingkungan EDA.

Risiko

- Boundary validity bbox tidak dapat diverifikasi.
- Operasi preprocessing umum (clip bbox, resize-aware augmentations) berpotensi menghasilkan error atau silent failure.
- Evaluasi model bisa bias jika bbox sebenarnya out-of-frame tanpa terdeteksi.

Hipotesis

- Metadata resolusi video berada di:
  - file manifest terpisah,
  - pipeline preprocessing video (belum dimuat ke EDA),
  - atau hanya tersedia pada level raw video (belum diekstrak).

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Validasi boundary **wajib** dilakukan sebelum:
  - clipping bbox,
  - resize / letterbox,
  - mosaic / mixup berbasis koordinat.
- Tanpa W/H, preprocessing berbasis geometri **tidak aman**.

Kaitan dengan Modeling

- Model detection/tracking akan mengasumsikan bbox berada dalam frame.
- Jika asumsi ini dilanggar secara tersembunyi, maka:
  - training loss bisa tidak stabil,
  - evaluasi mAP / MOT metrics bisa terdistorsi.

□ STEP B7 — Per-Video Label Coverage & Windowing Health

EDA_STEP_B7_Per_Video_Label_Cov…

---

1 □ Insight Detail per Sub-step / Output

□ Output B7.1 — Scope & Contract Confirmation

Fakta Penting

- Jumlah baris label dianalisis (haveVideo=True): **1,919,666**
- Jumlah video joinable: **961**
- Analisis hanya mencakup **frame berlabel**
- Tidak ada kolom penting yang hilang

Insight

- STEP B7 berjalan pada scope yang tepat untuk *label health*: hanya frame berlabel, tanpa asumsi negatif pada frame kosong.

Ini menunjukkan

- Semua metrik coverage yang dihitung merefleksikan **ketersediaan label aktual**, bukan artefak dari frame tanpa anotasi.

Resiko

- –

□ Output B7.2 — Per-Video Labeled Coverage Metrics

Fakta Penting

- Rata-rata `n_labeled_frames` per video: **195**
- Median `n_labeled_frames`: **202**
- Minimum `n_labeled_frames`: **77**
- `label_window_len` sangat dekat dengan `n_labeled_frames` pada mayoritas video

Insight

- Sebagian besar video memiliki **cakupan label yang hampir penuh** dalam window labelnya.

Ini menunjukkan

- Label windowing secara umum **padat dan konsisten**, bukan sporadis atau terputus-putus di dalam window.

Resiko

- –

□ Output B7.3 — Per-Video Label Density

Fakta Penting

- Rata-rata `mean_bboxes_per_labeled_frame`: **10.18**
- Median: **9.51**
- Rentang:
  - min: **1.65**
  - p90: **16.41**
  - p99: **23.26**
  - max: **25.72**

Insight

- Kepadatan objek per frame bervariasi cukup lebar antar video.

Ini menunjukkan

- Dataset mengandung campuran video dengan:
  - scene relatif sederhana (objek sedikit)
  - scene padat (banyak objek per frame)

Resiko

- –

□ Output B7.4 — Coverage & Window Length Distribution

Fakta Penting

- Rata-rata `coverage_ratio`: **0.991**
- Median `coverage_ratio`: **1.000**
- p10 `coverage_ratio`: **0.990**
- Minimum `coverage_ratio`: **0.381**

Insight

- Mayoritas video memiliki **coverage ratio mendekati 1**, artinya hampir semua frame dalam window memiliki label.

Ini menunjukkan

- Label windowing **sangat rapat dan seragam** pada sebagian besar video.
- Kasus coverage rendah adalah **outlier**, bukan pola umum.

Resiko

- –

□ Output B7.5 — Sparse & Extreme Video Candidates

Fakta Penting

- Threshold sparse: `n_labeled_frames ≤ 102`
- Proporsi video sparse: **2.60%** (≈25 video)
- Video paling sparse:
  - `n_labeled_frames` minimum: **77**
  - `coverage_ratio` minimum: **0.381**
- Video paling dense:
  - `mean_bboxes_per_labeled_frame` ≈ **25–26**

Insight

- Hanya sebagian kecil video yang benar-benar **label-sparse** atau **ekstrem dari sisi density**.

Ini menunjukkan

- Mayoritas dataset stabil, namun ada **subset kecil video** yang secara statistik berbeda dan layak ditandai.

Resiko

- –

□ Output B7.6 — Temporal Windowing Bias (Early Label Concentration)

Fakta Penting

- Distribusi `max_frameIndex_label`:
  - Median: **201**
  - p90–p99: **202–203**
  - Minimum: **81**
- Variasi relatif kecil antar video

Insight

- Label umumnya terkonsentrasi pada **rentang frame yang serupa** antar video.

Ini menunjukkan

- Tidak ada indikasi kuat bahwa sebagian besar video hanya dilabeli pada awal yang sangat pendek dibanding video lain.
- Bias temporal bersifat **terbatas dan terlokalisasi pada outlier**.

Resiko

- –

□ Output B7.7 — Label Coverage Severity Status

Fakta Penting

- Sparse video rate: **2.60%**
- Severity status: **minor**

Insight

- Prevalensi video label-sparse berada jauh di bawah ambang risiko.

Ini menunjukkan

- Dari perspektif *per-video label coverage*, dataset **layak dan stabil** untuk lanjut ke tahap berikutnya.

Resiko

- –

2️⃣ Insight Kesimpulan (Naratif — Terhubung)

Sub-step B7.1
Kontrak analisis ditegakkan dengan ketat: hanya frame berlabel yang dianalisis, sehingga semua metrik benar-benar merepresentasikan kesehatan label.

Sub-step B7.2–B7.4
Mayoritas video memiliki window label yang panjang dan padat, dengan coverage ratio mendekati 1. Ini menandakan bahwa label tidak hanya ada, tetapi **konsisten di dalam window**.

Sub-step B7.3 & B7.5
Variasi density bbox menunjukkan adanya perbedaan kompleksitas scene antar video, namun video ekstrem jumlahnya sangat terbatas.

Sub-step B7.6
Distribusi `max_frameIndex_label` yang sempit mengindikasikan bahwa label windowing relatif seragam dan tidak menunjukkan bias temporal sistemik.

Sub-step B7.7
Severity ditetapkan sebagai **minor**, menegaskan bahwa isu coverage hanya bersifat outlier-level.

3️⃣ Penjelasan Masalah yang Terjadi (Observasi, Risiko, Hipotesis)

Observasi

- Sebagian kecil video memiliki label yang lebih sedikit dan coverage ratio rendah.

Risiko

- Jika video-video ini diperlakukan sama tanpa penyesuaian, mereka dapat:
  - memberikan sinyal training yang lebih lemah,
  - atau mempengaruhi estimasi metrik pada level per-video.

Hipotesis

- Video label-sparse kemungkinan:
  - terpotong,

    - memiliki durasi pendek,
    - atau berasal dari kondisi pengambilan data yang berbeda.

---

4 Kaitan dengan Preprocessing dan Modeling

Kaitan dengan Preprocessing

- Video label-sparse (≈2.6%) dapat:
    - ditandai untuk audit manual,
    - diberi bobot lebih kecil,
    - atau dikeluarkan dari subset tertentu (opsional, bukan keharusan).
- Informasi density per video berguna untuk:
    - desain batching,
    - sampling yang lebih seimbang antar video.

Kaitan dengan Modeling

- Model perlu robust terhadap variasi density objek antar video.
- Strategi evaluasi sebaiknya:
    - mempertimbangkan metrik agregat,
    - dan, bila perlu, analisis terpisah untuk video ekstrem.

✅ Matriks 1 — Kumpulan Insight Informatif (Non-Problem)

Area	Insight Informatif Terpadu	Bukti Angka / Fakta Kunci	Implikasi Teknis (EDA-only)
Scope joinable	Dataset label terbagi jelas: total label 2,886,916 dan subset **joinable haveVideo=True = 1,919,666**	haveVideo=True: 1,919,666	Semua audit "MOT-ready" Block B valid jika dibatasi pada haveVideo=True
Kolom inti lengkap	Kontrak kolom inti MOT tersedia dan **tanpa missing** pada kolom kritikal	missing core = 0; required columns lengkap	Tidak perlu imputasi untuk kolom inti label pada subset joinable
Validitas bbox dasar	Tidak ada bbox invalid matematis (x1/y1 negatif, width/height ≤0)	all invalid checks = 0	Risiko error training karena bbox negatif/nol sangat rendah pada scope ini
Definisi kelas operasional	Himpunan kelas terobservasi stabil dan non-NaN; daftar kelas bisa didefinisikan langsung	11 kategori terobservasi; NaN category = 0	Mapping label→id bisa dibangun dari daftar kategori terobservasi
Duplikasi & konsistensi key	Tidak ada duplikasi literal, tidak ada konflik key logis (video, frame, id), dan **1 bbox per track per frame**	literal dup=0; logical key violation=0; multi-bbox per track/frame=0	Tracking annotation mengikuti kontrak MOT standar; integritas key kuat
Skala objek (global)	Distribusi area bbox long-tail: banyak objek kecil, dengan ekor objek besar	area: median 2,312 px²; p5≈297; p99≈164,061; max≈769,838	Karakteristik skala menjadi faktor penting untuk desain resolusi/augmentasi
Bentuk bbox (global)	Aspect ratio mayoritas moderat, namun ada ekor outlier bentuk ekstrem	AR median 1.2; p99 4.143; max 29.57	Model perlu robust terhadap variasi bentuk; outlier perlu awareness saat modeling
Kontinuitas tracking (global)	Track length bervariasi; mayoritas gap frame kecil dan jarang	track total 75,885; median len 15; gap>1 hanya 1.43%	Dataset cenderung punya tracking yang cukup konsisten secara agregat
Noise awal (definisi B5)	Tidak ada indikasi class-switch dalam track, dan tidak ada high IoU beda class pada sampel 2%	multi-class track=0; IoU≥0.7 beda class=0 (sample)	Indikasi awal label conflict (definisi B5) rendah pada scope joinable
Coverage per video	961 video joinable; window label umumnya padat dan seragam	mean labeled_frames 195; median 202; coverage_ratio mean 0.991; median 1.0	Label windowing relatif konsisten antar video; mendukung sampling yang lebih stabil

Area	Insight Informatif Terpadu	Bukti Angka / Fakta Kunci	Implikasi Teknis (EDA-only)
Kepadatan objek per frame	Density bbox per labeled frame bervariasi cukup lebar antar video	mean 10.18; median 9.51; max 25.72	Kompleksitas scene beragam; penting untuk batching/sampling strategy

⚠ Matriks 2 — Kumpulan Masalah (Observasi • Risiko • Hipotesis)

ID Masalah	Observasi	Risiko (jika tidak ditangani)	Hipotesis
B6 — [UNRESOLVED] Metadata W/H tidak tersedia	STEP B6 tidak menemukan metadata resolusi frame sama sekali (source none, coverage 0%), sehingga boundary check out-of-frame tidak bisa dilakukan; severity tercatat fatal di step B6	Validasi bbox terhadap batas frame tidak dapat diverifikasi → potensi **silent failure** saat preprocessing geometri (clip/resize/letterbox/mosaic) dan potensi bias evaluasi jika ada bbox out-of-frame yang tak terdeteksi	Metadata resolusi tersimpan terpisah (manifest/video info), belum dimuat ke environment EDA; atau hanya bisa diambil dari raw video (probe)
B3 — Dominasi small-object & long-tail scale	Area bbox sangat skewed; 5% terbawah area < ~297 px² dan 1% sangat kecil; ada outlier besar	Model bisa underperform pada small object; training bisa bias ke objek besar/menengah; augmentasi/resolusi input bisa tidak memadai	Dataset memang banyak objek kecil (dashcam jauh), atau label mencakup banyak objek distant; bukan error anotasi tetapi karakteristik distribusi
B3 — Outlier aspect ratio	Aspect ratio memiliki ekor panjang (p99 4.143, max 29.57) meski mayoritas moderat	Anchor/prior/assignment bisa tidak optimal; error prediksi pada objek sangat ramping/melebar; potensi instability jika outlier ekstrem tidak ditangani	Outlier berkaitan dengan kelas tertentu (mis. pedestrian/bicycle) atau kondisi visual (occlusion/truncation), atau bbox ketat pada objek memanjang
B4 — Track sangat pendek (fragmentasi ringan)	Track length ≤2 frame sebesar 6.70% dan ≤5 frame sebesar 19.66%	Jika dipakai tanpa strategi, track sangat pendek dapat melemahkan sinyal temporal untuk tracker/association; evaluasi ID-based bisa sensitif terhadap fragmentasi	Fragmentasi muncul karena occlusion, objek keluar frame cepat, atau label windowing/annotator behavior; bukan konflik key karena B2 bersih
B4 — Gap frame ekstrem walau jarang	Gap>1 hanya 1.43% tetapi max gap mencapai 169	Kasus gap ekstrem dapat menjadi sumber kesalahan association pada model tracking, dan memerlukan robust handling pada stage association/evaluation	Gap terjadi pada subset kecil track akibat hilangnya label pada beberapa frame atau objek hilang-lalu-muncul

ID Masalah	Observasi	Risiko (jika tidak ditangani)	Hipotesis
B5 — Metrik `num_class_changes` tidak informatif	B5.2 menyatakan 0 multi-class track, tetapi B5.3 menghasilkan nilai konstan 1.0 untuk semua track (indikasi definisi metrik baseline salah)	Jika metrik ini dipakai sebagai "signal" noise pada step lain, bisa menyesatkan interpretasi; merusak kredibilitas indikator	Implementasi menghitung perubahan pertama (prev NaN) sebagai "change", sehingga baseline menjadi 1 bukan 0
B7 — Video label-sparse (minor outliers)	~2.60% video (≈25) memiliki labeled frames rendah (threshold ≤102); minimum labeled_frames 77 dan minimum coverage_ratio 0.381	Video sparse dapat memberi sinyal training lebih lemah, dan bisa mengganggu fairness evaluasi per-video bila diperlakukan sama	Video tersebut kemungkinan berdurasi pendek, terpotong, atau berasal dari kondisi pengambilan berbeda
B1 — Class imbalance (long-tail)	Distribusi kategori sangat tidak seimbang: `car` dominan (1,473,500) sedangkan beberapa kelas kecil (`train` 1,053; `trailer` 854; `other person` 1,648)	Model cenderung bias ke kelas dominan; kelas long-tail rawan underfit/low recall; evaluasi mAP per class bisa timpang	Real-world dashcam bias (mobil dominan), dan definisi kelas long-tail memang jarang muncul