4) Dataset Cleaning Report（數據清理報告）

Date: 2025-12-24

Objective

- Normalize multiple FER datasets into a single **7-class canonical label space**:
  - `Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral`
- Produce cleaned dataset folders + unified CSV manifests.

Outputs (artifacts)

- Cleaning report: `Training_data_cleaned/clean_report.json`
- Cleaned manifests:
  - `Training_data_cleaned/classification_manifest.csv`
  - `Training_data_cleaned/classification_manifest_hq_train.csv`
  - `Training_data_cleaned/expw_full_manifest.csv`, `Training_data_cleaned/expw_hq_manifest.csv`
- Manifest validation reports:
  - `outputs/manifest_validation.json`
  - `outputs/manifest_validation_all_with_expw.json`

Cleaning method summary

Based on `Training_data_cleaned/clean_report.json`:

- Mode: `link` (creates cleaned structure via links)
- Per-dataset adapters:
  - Folder datasets (AffectNet balanced, FER2013 uniform 7, FERPlus, RAF-DB basic)
  - YOLO-format AffectNet: class-name mapping + dropping unwanted label lines
  - RAF compound: mapping to canonical subset

Canonical class mapping notes

- FERPlus includes an extra class `contempt` that is excluded during cleaning.

- AffectNet YOLO format mapping drops the original “Contempt” class.

Validation results (no missing paths)

From `outputs/manifest_validation_all_with_expw.json`:

- Manifest: `Training_data_cleaned/classification_manifest.csv`
- Total rows: **466,284**
- Valid rows: **466,284**
- Missing paths: **0**
- Bad labels: **0**
- Decode sample: attempted **300**, ok **300**, fail **0**

Source composition (rows)

From `outputs/manifest_validation_all_with_expw.json` `counts.by_source`:

- `fer2013_uniform_7`: **140,000**
- `ferplus`: **138,526**
- `expw_full`: **91,793**
- `affectnet_full_balanced`: **71,764**
- `rafdb_basic`: **15,339**
- `rafml_argmax`: **4,908**
- `rafdb_compound_mapped`: **3,954**

Quality gates applied

- Canonical label enforcement (7 classes only)
- Path existence check for every CSV row
- Random decode sampling to catch broken images

Next steps

- If needed: add stronger image-level QA (blur/occlusion thresholds) and re-export `*_hq_*.csv` manifests.