VIEW POINT

FROM DATA VOLUME TO DATA VALUE: CURATING ENTERPRISE COMPUTER VISION DATASETS THROUGH ACTIVE LEARNING AND HUMAN-IN-THE-LOOP GOVERNANCE

Abstract

Modern computer vision systems increasingly face a data-quality challenge rather than a model-capacity challenge. Large enterprise image and video datasets often contain redundant, low-quality, duplicated or inconsistently labeled samples that increase annotation cost while contributing limited incremental training value. This white paper presents a human-in-the-loop dataset curation framework designed to transform large-scale enterprise image collections into efficient, governed and reusable training datasets. The framework combines automated image quality filtering, AI-assisted pre-labeling, active-learning-based sample prioritization, confidence-driven human validation, dataset lineage and feedback-driven improvement loops. It operationalizes annotation not as a one-time labeling task, but as a continuous data curation lifecycle. Aggregated implementation observations from anonymized enterprise computer vision workflows indicate up to 71% reduction in bounding-box annotation handling time, up to 80% reduction in segmentation handling time, approximately 60% reduction in turnaround time and approximately 45% overall operational efficiency uplift in selected workflows.

Infosys
Navigate your next

1. Business context: Why data value matters

Enterprise computer vision teams are collecting more images, more video frames and more visual evidence than ever before. However, a larger data lake does not automatically create better AI outcomes. Poor-quality images, duplicate samples, inconsistent annotation rules and non-versioned labels can slow down model development and reduce trust in downstream outputs.

For scalable AI programs, the operating question must change from "how much data can be labeled?" to "which data creates the highest training value?" This shift requires an integrated curation model that qualifies input data, prioritizes valuable samples, uses automation to accelerate repeatable annotation work, retains human judgment for exceptions and converts approved labels into governed reusable assets.

Core thesis Efficient learning is not achieved only by adding more labeled data. It is achieved by selecting, improving, governing and reusing the right data with targeted human oversight.

2. Enterprise dataset curation challenges

Enterprise computer vision programs often fail to realize full value from available visual data because the issue is not data availability alone, but data usability.

The recurring challenges are less about collecting visual data and more about converting the right visual data into reusable training assets.

- **Low-value input noise:** Raw collections can include blurred, dark, corrupted, duplicated or irrelevant images, consuming effort without adding meaningful training value.
- **Manual effort leakage:** Teams may label every available image instead of prioritizing high-value samples, which increases cost and turnaround while limiting model impact.
- **Quality variation:** High-volume work can create label drift across annotators, tools, batches and evolving guidelines, reducing training and validation reliability.
- **Limited dataset reuse:** Approved labels are not always versioned, catalogued or governed as reusable assets, forcing similar use cases to restart from raw inputs.
- **Weak feedback loops:** Model errors, drift signals and new edge cases do not consistently return to the annotation pipeline, leaving the dataset static as conditions evolve.

3. Differentiation from current industry approaches

Most industry approaches improve one part of the annotation chain — workforce scale, labeling tooling, active learning or data storage. These capabilities are useful, but they often remain fragmented across teams and platforms. The proposed approach is different in that it treats annotation as a governed dataset value lifecycle rather than a task-level labeling operation.

Industry approach	Typical focus	Limitation	Proposed advantage
Manual annotation factory	High-volume label production through workforce scale.	Effort scales with data volume and may include redundant or low-value samples.	Human effort is reserved for uncertain, high-risk and high-value samples.
AI-assisted labeling tool	Pre-labeling and productivity acceleration.	Speed improves, but reusable governed datasets may not be created.	Pre-labeling is combined with lineage, validation, versioning and reuse controls.
Active learning platform	Prioritizing samples for annotation or review.	Sample selection can be disconnected from delivery workflows and governance	Active learning is embedded into the end-to-end annotation operating model.
Dataset repository	Storage and discoverability of data assets.	Repositories may not drive feedback-based refresh or approval governance.	Bronze, Silver and Gold dataset states enable controlled reuse and refresh.
Traditional outsourcing model	Capacity, SLA and throughput management.	Success is often measured by volume rather than model-readiness.	Success is measured by data value, quality confidence, reuse and model-improvement readiness.

What makes this framework different

The framework connects sample selection, AI-generated labels, human validation, dataset approval, lineage and model feedback into one reusable operating model. Human expertise becomes a governance lever and the dataset becomes a strategic asset, not merely an output of labeling work.

4. Proposed framework

The proposed framework treats annotation as a continuous dataset curation lifecycle. It combines automated curation agents, active learning signals, human validation gates and dataset governance. The framework progresses raw visual data through qualification, machine assistance, targeted review and approved reusable dataset states.

4.1 Framework design principles

The design is anchored on five principles: Data value first, automation by default, human-in-the-loop governance, continuous learning and reusable data assets. Together, these principles shift the operating model from labeling everything to curating what matters. Human reviewers remain central, but their role moves from repetitive first pass labeling to validation, adjudication, exception handling and final quality governance.

4.2 End-to-end framework flow

The lifecycle below is now positioned inside the main framework section rather than before the abstract, so the reader first understands the business context and then sees how the operating model works.

Human-in-the-Loop Dataset Curation Lifecycle

From raw enterprise images to governed, reusable gold datasets

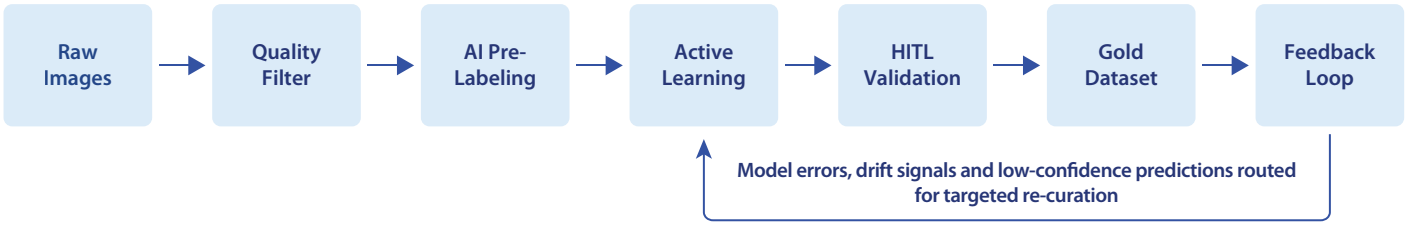

Figure 1. Human-in-the-loop dataset curation lifecycle for efficient learning.

The flow starts with raw images, applies quality filters, generates AI pre-labels, prioritizes samples through active learning, validates outputs through human-in-the-loop review, promotes approved labels into gold datasets and sends model errors or low-confidence predictions back into the loop for targeted refresh.

Data volume to data value stack

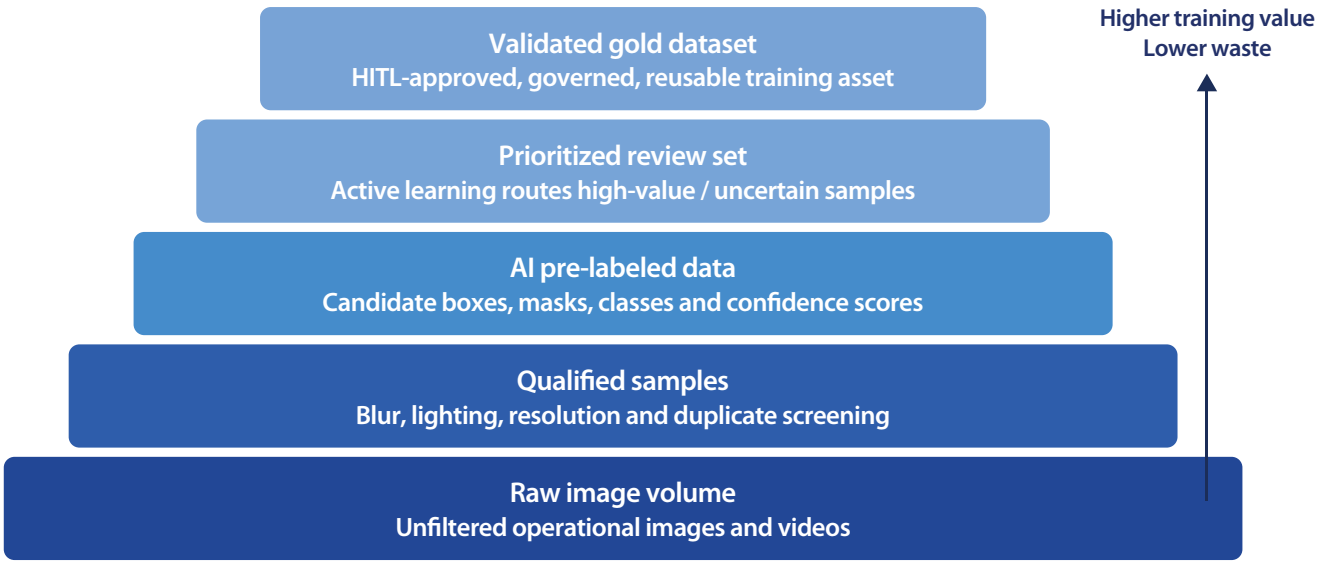

Figure 2. Data Volume to Data Value stack showing progressive curation and governance.

Figure 3. Agent-assisted annotation workflow across image quality, bounding box and segmentation stages

Operationally, the framework can be read as a seven-stage flow. Raw visual data is first registered with source and batch metadata. Automated data qualification screens are given as input for usability. AI-assisted pre-labeling generates candidate boxes, masks, classes and confidence scores. Active learning prioritizes uncertain, diverse or high-impact samples. Human reviewers then approve, correct, reject or escalate outputs. Validated records are promoted into governed dataset states, and model feedback continuously refreshes the curation queue.

4.3 Reference operating layers

For enterprise implementation, the framework works through five operating layers. The data intake and observability layer captures source, batch and metadata traceability. The curation intelligence layer screens quality, detects duplication, and ranks samples. The annotation automation layer generates candidate labels for bounding boxes, segmentation masks, classes and metadata. The human validation layer manages review, correction and exception handling. Finally, the dataset governance layer maintains lineage, schema, version states, quality scores and reuse controls.

5. Sample input and output

The following sample is retained as a table because it is easier to read as a side-by-side technical example and remains fully illustrative and anonymized.

Sample input: raw image metadata


```
{
"image_id": "IMG_0001",
"source": "enterprise_visual_workflow",
"capture_type": "image",
"quality_score": 0.84,
"duplicate_flag": false,
"use_case": "object_localization"
}
```


Sample output: curated annotation record


```
{
 "dataset_state": "Gold",
"image_id": "IMG_0001",
"label_type": "bounding_box",
"class": "target_object",
"coordinates": [124, 88, 302, 241],
"model_confidence": 0.91,
"human_action": "validated_no_edit",
"schema_version": "v1.0",
"reuse_status": "approved"
}
```


6. Implementation overview

The framework was applied across enterprise computer vision workflows involving object localization, segmentation and visual inspection. The previous operating model required annotators to open each image, identify relevant objects, manually draw bounding boxes or polygons, assign classes, correct coordinates and submit outputs for review. The updated workflow introduced automated image qualification, machine-generated candidate labels, confidence-based work routing and human validation before dataset approval.

A dataset governance layer was added to make the outputs reusable. Approved annotations were organized with normalized files, quality metadata, lineage, schema version and dataset state. This enabled future model training, validation and refresh cycles to begin from curated assets rather than restarting from raw data each time.

7. Observed outcomes

The following metrics are aggregated and anonymized. This table is retained because it provides a compact evidence view of the operational impact.

Metric	Baseline / before	After curation framework	Observed improvement
Bounding-box annotation handling time	Approx. 7 minutes per image	Approx. 2 minutes per image	Up to 71% reduction
Segmentation handling time	Approx. 15 minutes per image	Approx. 3 minutes per image	Up to 80% reduction
End-to-end turnaround time	Baseline manual workflow	Automation-assisted workflow with HITL validation	Approx. 60% reduction
Overall operational efficiency	Manual annotation-heavy model	Curated workflow with pre-labeling and targeted review	Approx. 45% uplift
Human validation pass rate	Not applicable in manual baseline	Selected AI-generated outputs accepted with no edits	Approx. 30% no-edit validation pass
Manual effort intensity	High-volume manual labeling and review	Human review focused on low-confidence cases and exceptions	Approx. 40–70% reduction range in selected workflows

Aggregated Implementation Impact Metrics

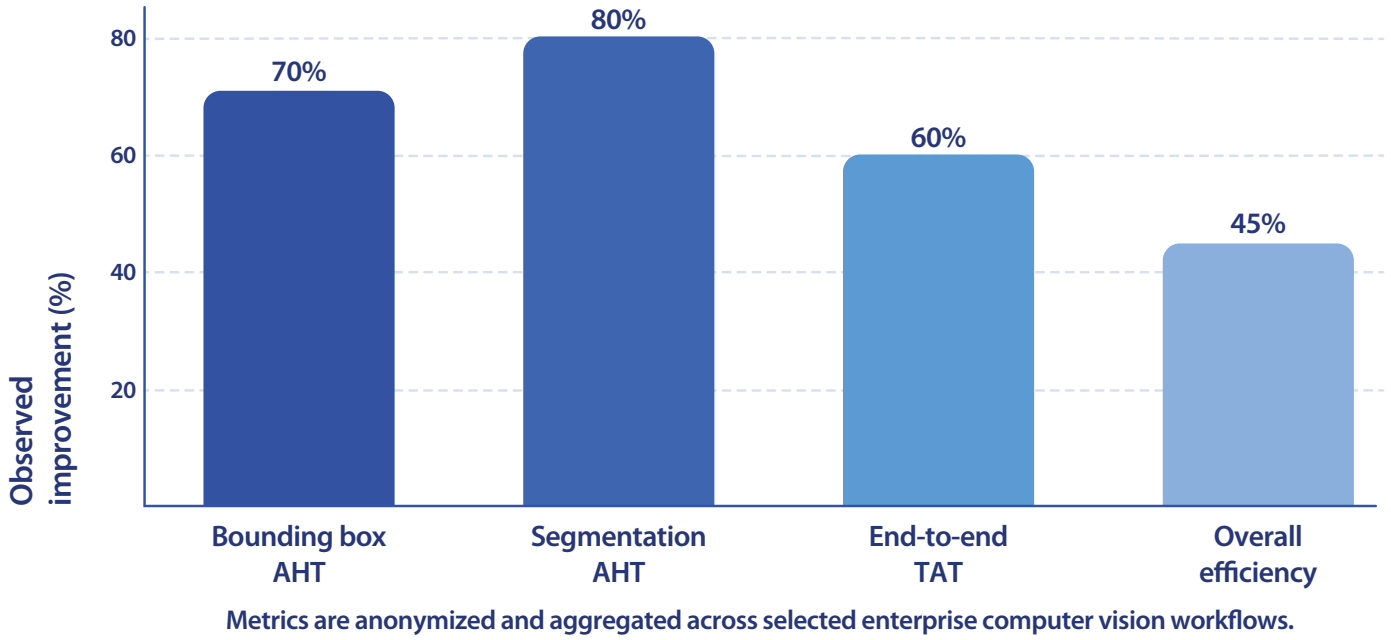

Figure 4. Aggregated implementation impact metrics from anonymized enterprise workflows.

8. How the framework improves data value

Data value is created when each step in the curation lifecycle removes waste, improves label reliability and increases the reuse potential of approved datasets. Quality filtering prevents unusable images from consuming annotation effort. AI-assisted pre-labeling shifts effort away from first-pass drawing toward validation and correction. Active learning directs reviewers to low-confidence, high-impact or diverse samples where human expertise is most valuable. Human-in-the-loop validation maintains quality governance, while dataset lineage and versioning convert one-time outputs into reusable enterprise training assets. Finally, feedback-driven refresh routes model errors, drift signals and new edge cases back into the curation queue to support continuous improvement.

9. Operating model and governance considerations

For enterprise adoption, the framework should be managed as an operating model rather than a standalone automation workflow. Dataset lineage should track source, batch, schema version and transformation steps. Quality gates should define pass/fail checks for input quality, label quality and reviewer approval. Confidence thresholds should route low-confidence predictions to human review, while version control should promote datasets through Bronze, Silver and Gold states. Feedback management should keep production errors and edge cases connected to the curation queue, creating a closed-loop improvement cycle.

10. Practical use cases

The curation approach is applicable wherever visual data must be converted into high-confidence training assets. In retail and supply chain vision, the framework supports object detection, shelf intelligence, asset condition review and item-level visual validation. In industrial inspection, it supports defect detection, component localization and equipment review. In geospatial and mapping operations, it supports image-based feature extraction, road assets, building footprints, change detection and map enrichment. In transportation and logistics, it supports vehicles, infrastructure, route assets, and warehouse image annotation. It can also support enterprise visual knowledge bases by reducing duplicate labeling and increasing reuse of governed training data.

11. Conclusion

Enterprise AI training data operations must evolve from volume-centric annotation toward value-centric dataset curation. The proposed framework combines image qualification, AI-assisted pre-labeling, active learning, human-in-the-loop validation and dataset governance to convert raw visual data into reusable gold datasets. The observed improvements in handling time, turnaround time and operational efficiency show that efficient learning can be advanced by treating annotation as a governed, feedback-driven data lifecycle rather than a one-time labor-intensive activity.

Authors

Karthigaiselvan R
Head – GIS and AI Training Services

Karthi heads GIS and AI annotation services and has 22+ years of techno-commercial experience in GIS and AI training services. He specializes in building innovative solutions and ensuring seamless project delivery. His expertise spans solution design, execution, and driving business growth. He did his M. Tech in remote sensing from Anna University, Chennai, and his engineering in electrical & electronics from Government College of Technology, Coimbatore.

Shyam Rao
Head – Digital Business Services

Shyam heads the Digital Business Services Unit at Infosys BPO and has over 21 years of experience in global business service delivery with specific focus on Digital Marketing Services (DCX) and GIS operations. Shyam is responsible for marketing, sales and delivery of digital operations for leading Fortune 500 clients and has led several consulting engagements to define the digital operating model and drive innovation in their digital operations. He is an alumnus of London School of Economics and the Digital Marketing Institute of Ireland.