Fine-Grained Visual Understanding for Multimodal and Trustworthy AI

Syuan-Hao Li, Yi-Jui Huang, Lei Wang, Yongsheng Gao

Griffith University | NYCU | Data61/CSIRO

WWW Companion 2026 | Dubai, UAE

What Is Fine-Grained Visual Understanding?

Goal: distinguish visually similar but semantically different objects

Input: image / text / metadata / sensor signals
Output: evidence-grounded decision
Uses: biodiversity, agriculture, medicine, retail, safety

Key idea: the decisive cue is subtle, localized, and domain-specific.

Motivation and Research Questions

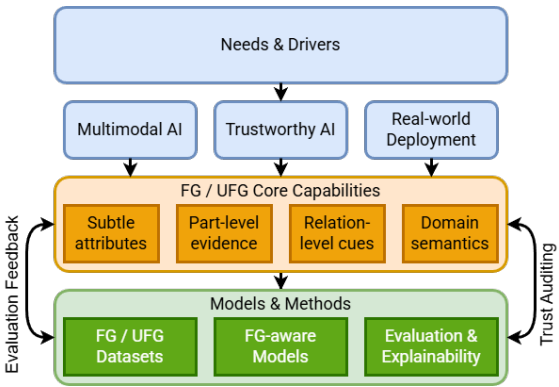
The problem is becoming more important

- Multimodal models need part- and attribute-level grounding
- Trust failures often hide inside fine-grained slices
- Real deployments require expert-level distinctions
- Existing benchmarks still reward shortcuts

We ask three questions:

Why does FG/UFG matter for multimodal AI?
How do datasets and methods evolve?
What makes FG/UFG trustworthy?

Our Study Framework

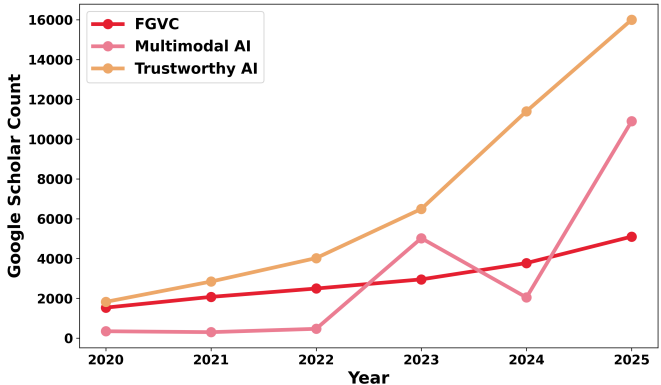


Source: Li et al., WWW Companion 2026.

We analyze:

- FG / UFG datasets
- FG-aware model families
- Evaluation and explainability
- Multimodal alignment
- Trustworthy deployment

Why FG / UFG Matters Now

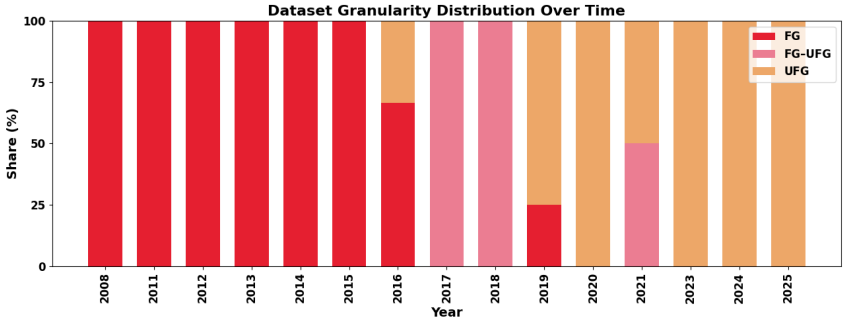


FGVC grows steadily

Multimodal AI and trustworthy AI surge rapidly

Google Scholar trend analysis reported in Li et al., 2026.

Dataset Trends

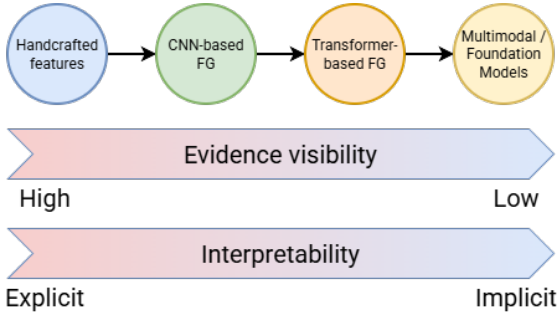


Earlier datasets: classical FG categories

Recent datasets: UFG, expert-defined, long-tailed

Representative dataset inventory from Li et al., 2026.

Main Model Trend



From explicit evidence

- SIFT / HOG / BoW
- Manual parts and keypoints

To implicit evidence

- CNNs and transformers
- CLIP-style multimodal models

Multimodal and Trustworthy AI

FG/UFG acts as semantic anchors

- Align visual parts with language phrases and metadata
- Reduce hallucination by checking evidence-level claims
- Reveal bias under attributes, subtypes, and rare slices
- Support explanations through parts, attributes, and relations

Takeaway: trust depends on whether the model used the right evidence.

Practical Guidelines & Future

Better evaluation should include:

- Accuracy plus evidence localization
- Long-tail and distribution-shift testing
- Attribute- and part-level bias audits
- Ambiguity, uncertainty, and expert disagreement
- Human evaluation for high-stakes use

Future directions:

FG-aware representation learning, data-efficient UFG, expert-in-the-loop learning, temporal reasoning, auditable foundation models.

Conclusion

Main message:

Fine-grained visual understanding is moving from a niche classification benchmark to a foundation for grounded, multimodal, and trustworthy AI.

Contributions:

- Dataset and method taxonomy
- Trend analysis for FGVC, multimodal AI, and trustworthy AI
- Trustworthiness view: bias, robustness, explainability
- Research roadmap for FG/UFG multimodal systems

Thank you! **Questions & answers**