Edge-based Cross-modal Semantic Routed Retrieval

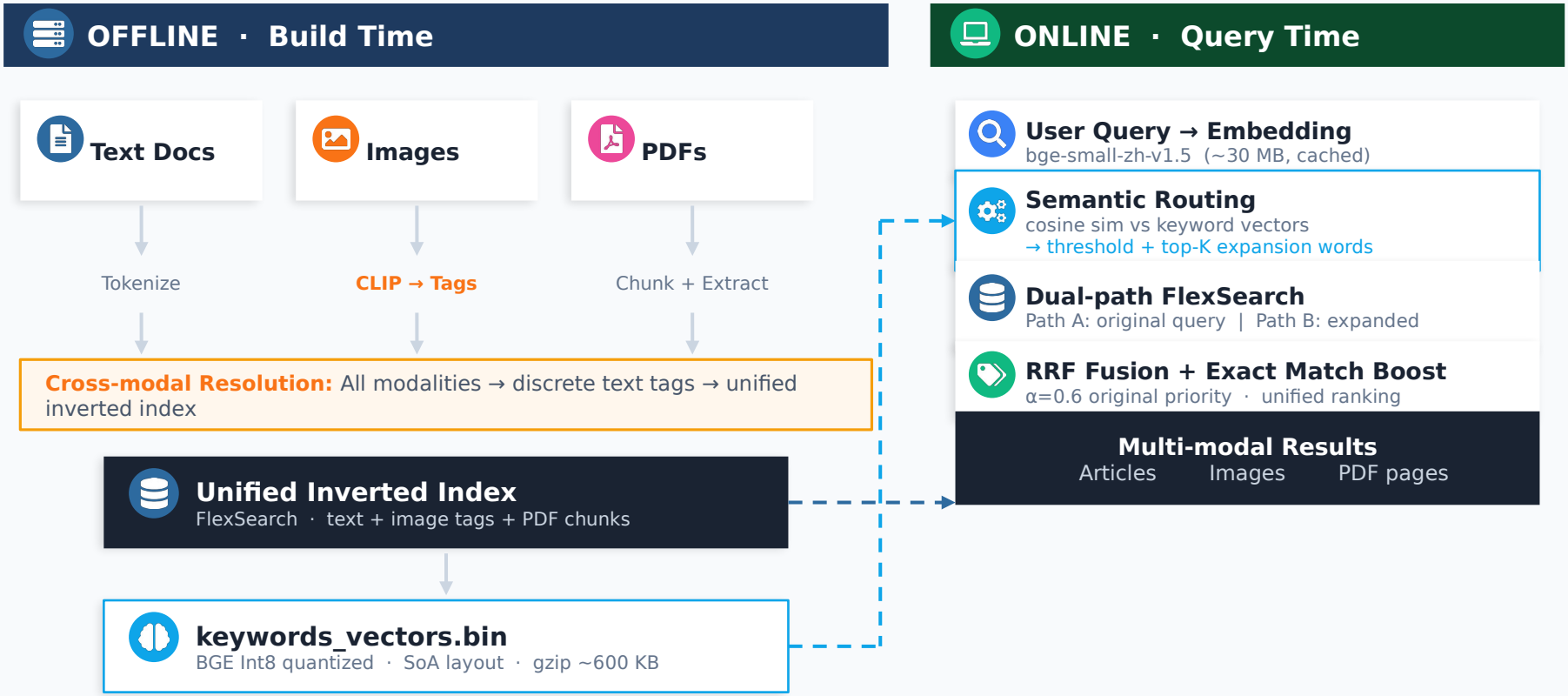
Resolving cross-modal search at build time, not query time.

How can we design a retrieval architecture that achieves semantic-level search quality on edge devices while maintaining explainability and sub-200ms latency?

Yuxu Ge · MSc Artificial Intelligence, University of York · https://Yuxu.Ge

February 2026

System Architecture



Key Contributions

Word-level Semantic Routing

Reduce online retrieval from O(n×d) document vectors to O(k×d) keyword vectors. k ≈ 3000 vs n ≈ 10,000+.

Cross-modal Offline Resolution

CLIP auto-tagging converts images to discrete labels at build time. Online search sees only text — zero cross-modal computation at query time.

Progressive Enhancement

Search works immediately with keyword matching (~1s load). Semantic routing activates silently once the 30 MB model is cached.

Edge-native & Scalable

Fully on-device, no server needed. Int8 quantization + SoA layout defer sharding up to 40K terms. All data stays local.