VFR Technical Report 2026-07 · 14 October 2026

Segment-Level Late Interaction for Moment Retrieval at Billion Scale

Aino Lehtovaara, Mikko Rantasalo, Sanna Kivimäki, Tomasz Wierzba
VideoFinder Research — Helsinki · Kadrowa Labs Nordics Oy, 00180 Helsinki

Abstract

Dense moment retrieval systems usually compress each candidate clip into a single vector, which discards the temporal structure that users search for. We describe the retrieval stack behind VideoFinder moment search, which indexes 1.1 billion half-second segments with 64-byte product-quantised vectors and applies a lightweight late-interaction re-ranker only to the top segments per shard. A temporal merge step converts segment hits into contiguous moments before clip-level ranking. On HETKI-40K the system improves R@1 at IoU 0.5 from 38.2 to 46.9 over a single-vector baseline while keeping median server latency at 38 ms on CPU-only search nodes. We report the engineering trade-offs that mattered most: shard fan-out, quantiser retraining cadence, and how often the merge step should be allowed to bridge short gaps.

About this document

This is the public extended abstract. The full report, including appendices, ablations and evaluation scripts, is available on request from research@videofinder.link. Code and data referenced in this work are listed at https://helsinki.videofinder.link/open-source.html and /datasets.html.

Cite as

```
@techreport{lehtovaara2026segmentlevel,
  title = {Segment-Level Late Interaction for Moment Retrieval at Billion
  Scale},
  institution = {VideoFinder Research, Kadrowa Labs Nordics Oy},
  year = {2026}, number = {VFR-TR-2026-07}
}
```