VFR Technical Report 2026-02 · 30 March 2026

Distilling Video–Text Encoders for CPU-Only Indexing

Priya Raghunathan, Sanna Kivimäki, Ville Suominen
VideoFinder Research — Helsinki · Kadrowa Labs Nordics Oy, 00180 Helsinki

Abstract

Indexing a growing stock library with a 1.1-billion-parameter video–text encoder requires a GPU fleet whose cost grows linearly with uploads. We distil the encoder into an 86-million-parameter student that runs at 41 segments per second per CPU core after int8 quantisation. Naive feature distillation preserves object semantics but loses temporal information; we show that adding a relational loss over segment pairs within the same shot and a short-range temporal contrastive term recovers most of the gap. The student retains 96.1% of the teacher's R@5 on HETKI-40K while reducing indexing cost per clip by 14×. We describe the failure modes we encountered, including calibration drift after quantisation and the effect of shot boundaries on pooled features.

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{raghunathan2026distilling,
  title = {Distilling Video–Text Encoders for CPU-Only Indexing},
  institution = {VideoFinder Research, Kadrowa Labs Nordics Oy},
  year = {2026}, number = {VFR-TR-2026-02}
}
```