Preprint · 27 August 2026

Motion Tokens: Disentangling Camera and Subject Motion from Dense Optical Flow

Elias Nordqvist, Sanna Kivimäki, Hanna Øvergaard
VideoFinder Research — Helsinki · Kadrowa Labs Nordics Oy, 00180 Helsinki

Abstract

Text queries about video frequently describe motion (“slow push-in”, “drifts upward”, “whip pan to the left”), yet image–text encoders applied frame by frame are nearly blind to it. We introduce motion tokens: a compact set of 16 learned descriptors per segment derived from dense optical flow, trained with a contrastive objective against motion phrases mined from 2.3 million editor-written captions. A homography-based decomposition separates global camera motion from residual subject motion before tokenisation. Adding motion tokens to the segment embedding improves retrieval on motion-centric HETKI queries by 11.4 points R@5 and reduces confusion between opposite camera moves by 63%, at a cost of 0.9 ms per segment at indexing time.

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{nordqvist2026motion,
  title = {Motion Tokens: Disentangling Camera and Subject Motion from Dense
  Optical Flow},
  institution = {VideoFinder Research, Kadrowa Labs Nordics Oy},
  year = {2026}, number = {VFR-TR-2026-05}
}
```