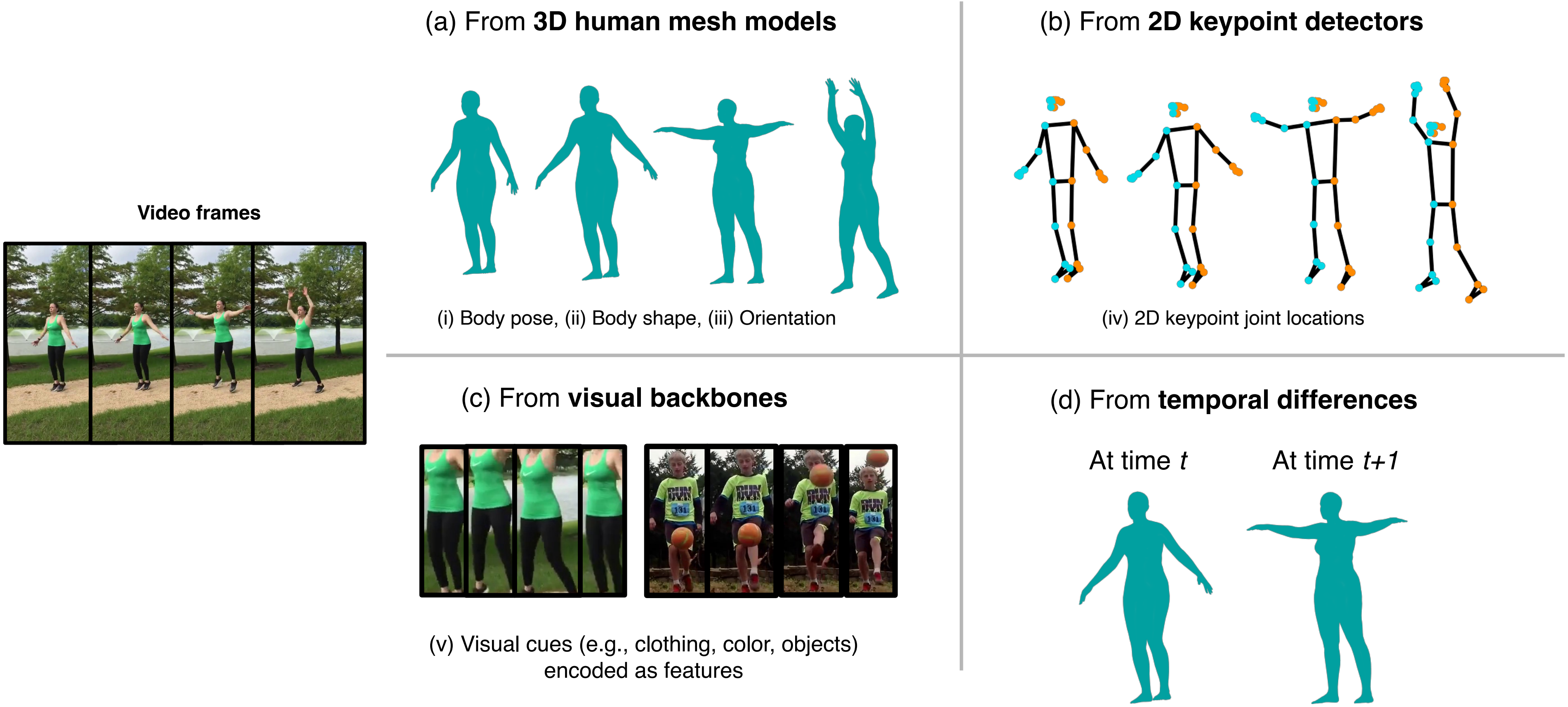
Video frames
(a) From 3D human mesh models
(i) Body pose, (ii) Body shape, (iii) Orientation
(b) From 2D keypoint detectors
(iv) 2D keypoint joint locations
(c) From visual backbones
(v) Visual cues (e.g., clothing, color, objects) encoded as features
(d) From temporal differences
At time t
At time t+1