NEURAL INFORMATION PROCESSING SYSTEMS

VGG

UNIVERSITY OF OXFORD

Chirality in Action: Time-aware Video Representation Learning by Latent Straightening

NeurIPS 2025

Piyush Bagad

Prof. Andrew Zisserman

Distinguishing temporal change in a video

Flavour 1: Arrow of time

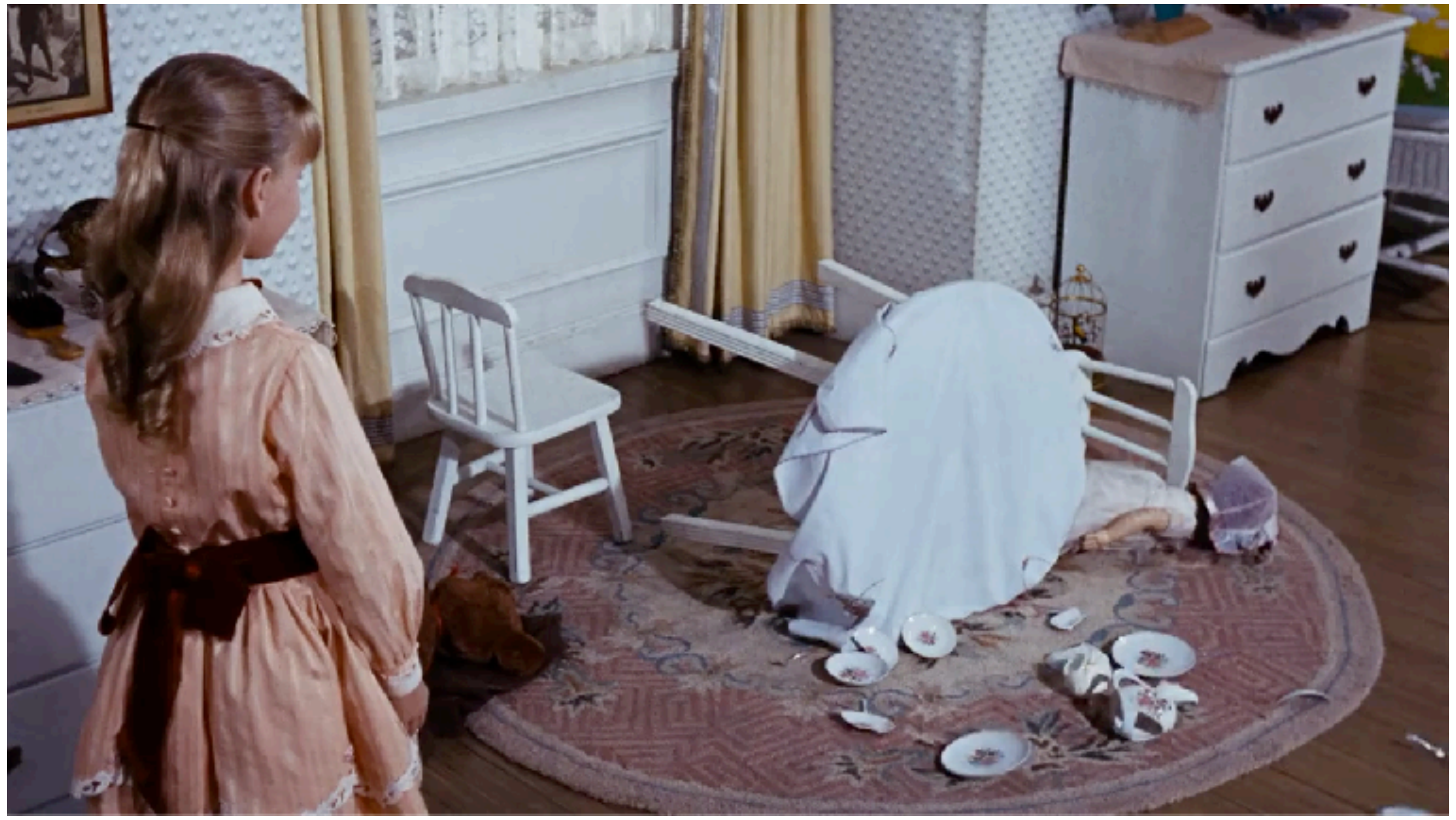
- Distinguish between “forward” and “reverse” videos

Learning and Using the Arrow of Time,
Donglai Wei, Joseph Lim, Andrew Zisserman and William T. Freeman, CVPR 2018.

Distinguishing temporal change in a video

Flavour 1: Arrow of time

- Distinguish between “forward” and “reverse” videos

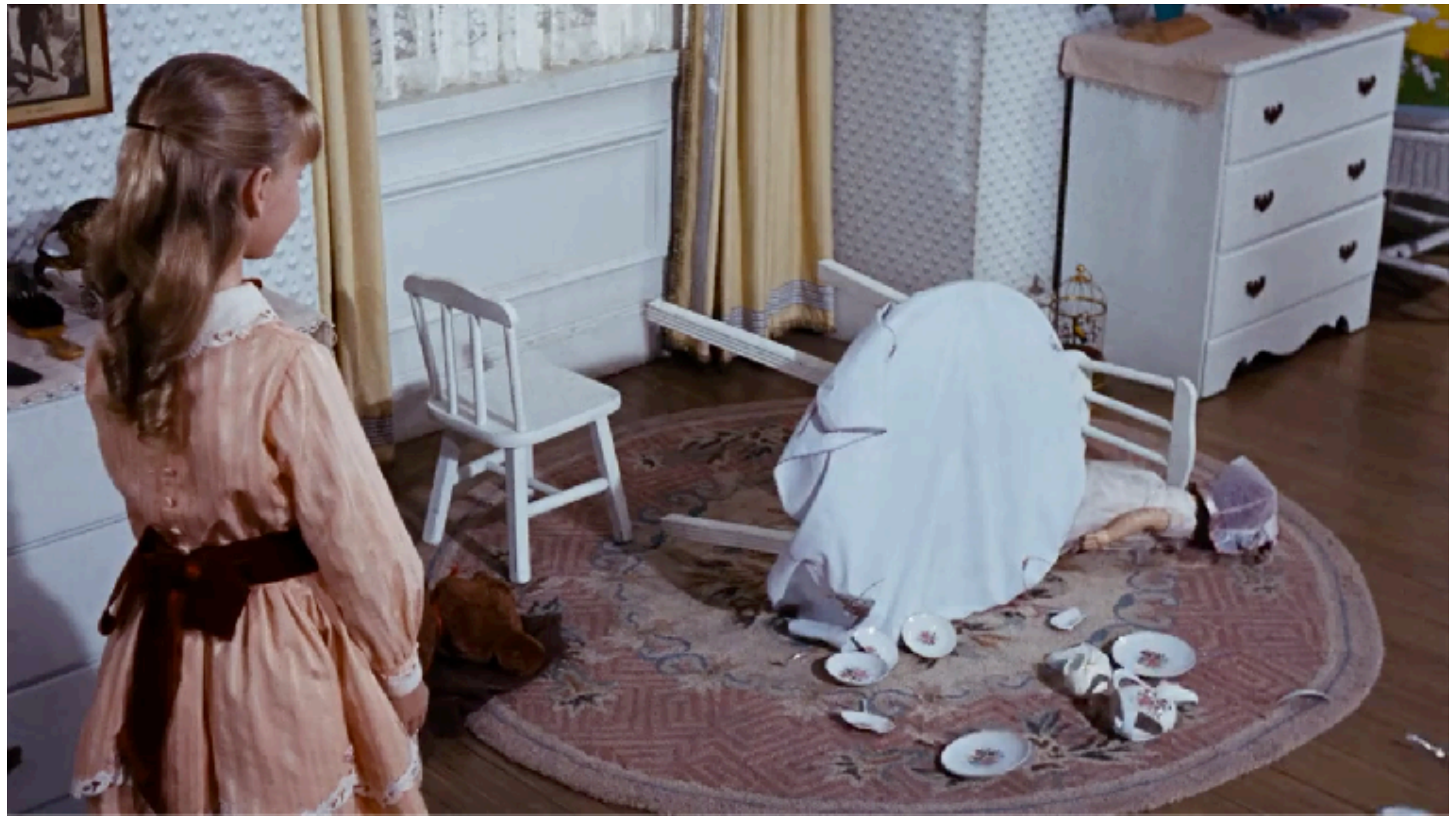


Learning and Using the Arrow of Time,
Donglai Wei, Joseph Lim, Andrew Zisserman and William T. Freeman, CVPR 2018.

Distinguishing temporal change in a video

Flavour 1: Arrow of time

- Distinguish between "forward" and "reverse" videos



- Cue: Reversed videos are often **physically implausible**

Learning and Using the Arrow of Time,
Donglai Wei, Joseph Lim, Andrew Zisserman and William T. Freeman, CVPR 2018.

Distinguishing temporal change in a video

Flavour 2: Temporally opposite (*chiral*) actions

Distinguishing temporal change in a video

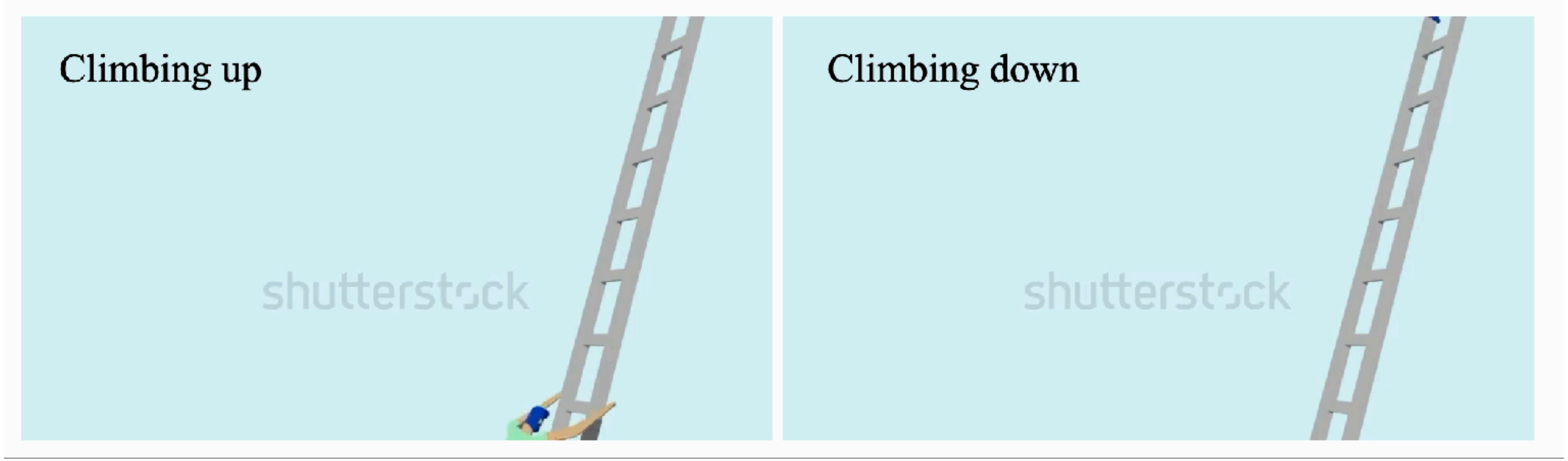
Flavour 2: Temporally opposite (*chiral*) actions

- Such actions have spatially similar contexts but temporally opposite verbs

Distinguishing temporal change in a video

Flavour 2: Temporally opposite (*chiral*) actions

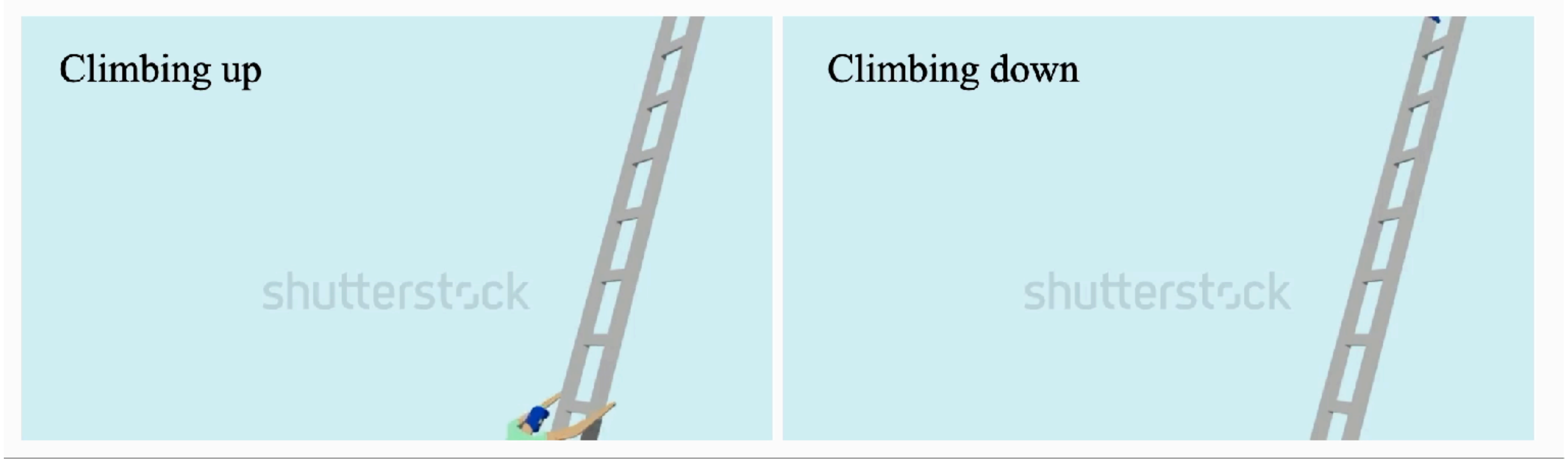
- Such actions have spatially similar contexts but temporally opposite verbs



Distinguishing temporal change in a video

Flavour 2: Temporally opposite (*chiral*) actions

- Such actions have spatially similar contexts but temporally opposite verbs



- Cue: **Visual change** (e.g., change in position)

(The lack of) time in video representations

Prior work

- A vast majority of the video benchmarks do not test for time-awareness
  - *Can be solved with a single frame or few frames without temporal modelling*

(The lack of) time in video representations

Prior work

- A vast majority of the video benchmarks do not test for time-awareness
  - *Can be solved with a single frame or few frames without temporal modelling*
- Many contemporary methods do not explicitly model temporal change
  - *E.g., Perception Encoder uses average pool over frame embeddings*

(The lack of) time in video representations

Prior work

- A vast majority of the video benchmarks do not test for time-awareness
  - *Can be solved with a single frame or few frames without temporal modelling*
- Many contemporary methods do not explicitly model temporal change
  - *E.g., Perception Encoder uses average pool over frame embeddings*
- Native video models like V-JEPA jointly model space-time but are very expensive to train from scratch

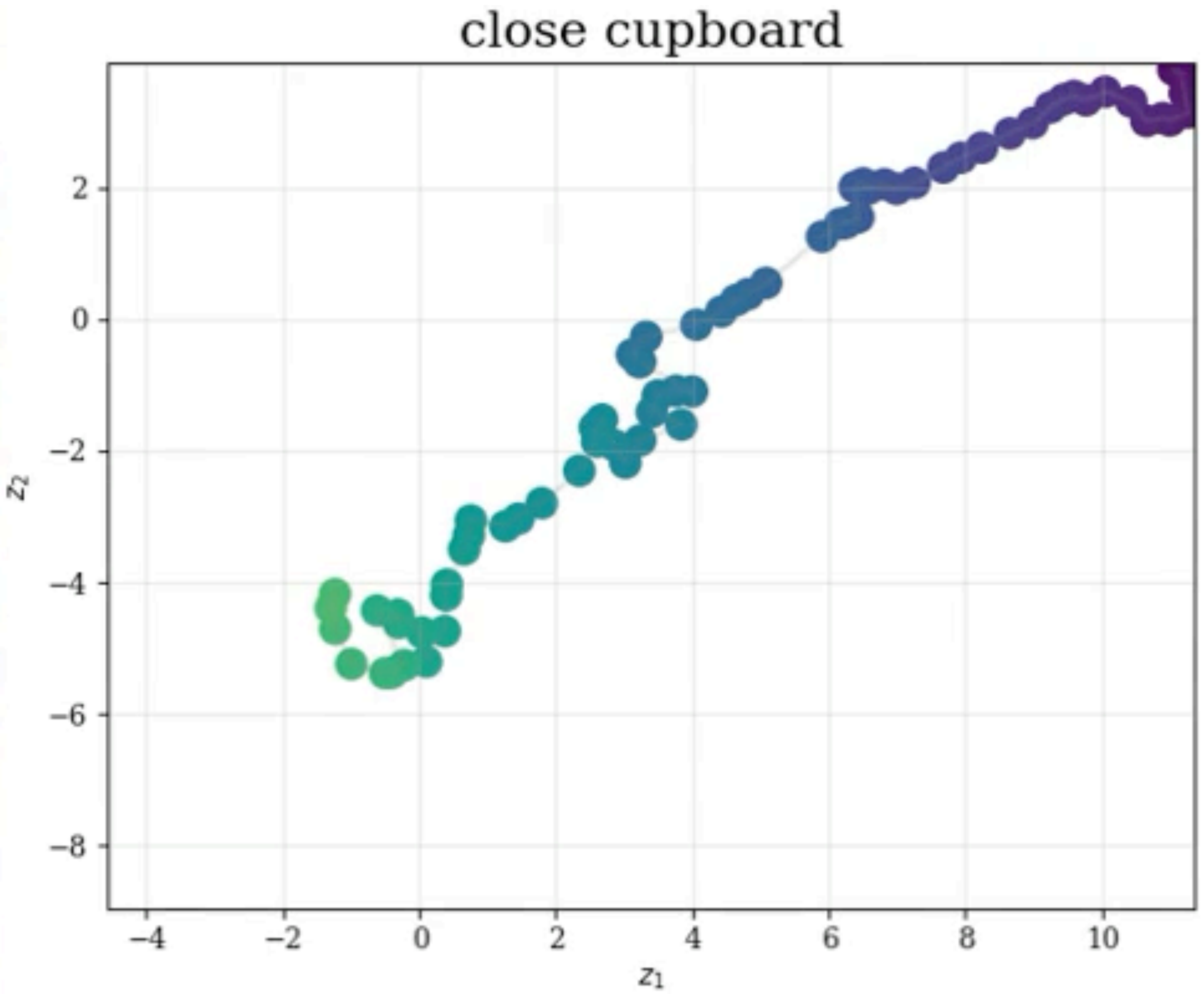
Introducing time in video representations

Outline of our work

1. A time-aware, compact video embedding model
2. A benchmark and measure of time-sensitivity (based on *chiral* actions)
3. Experimental evaluation

Building intuition

tSNE embeddings of per-frame DINOv2 features for a video



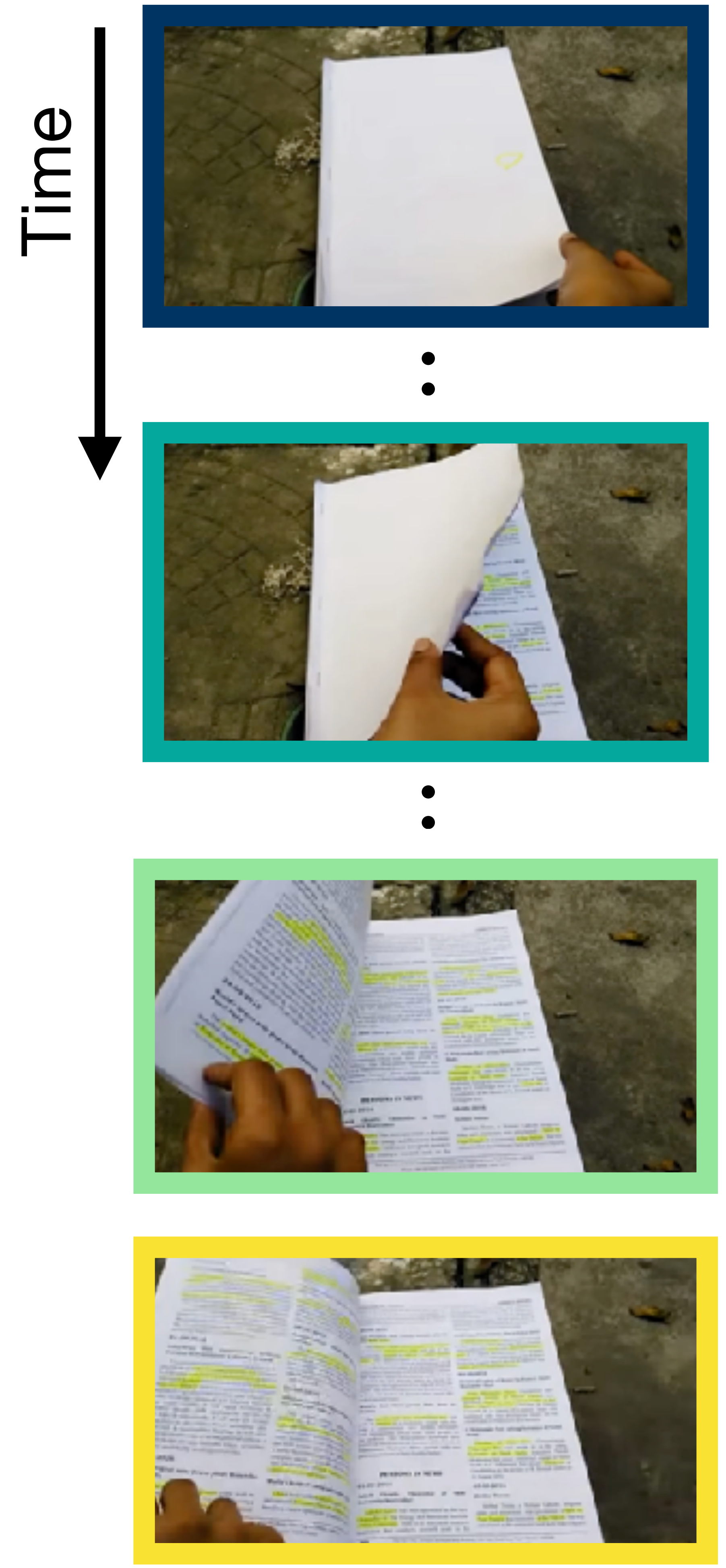
From intuition to the model

Key observation: The sequence of per-frame DINOv2 features lie on a time-sensitive trajectory!

If we can learn to "summarise" this trajectory in a single vector, then we have a time-sensitive embedding.

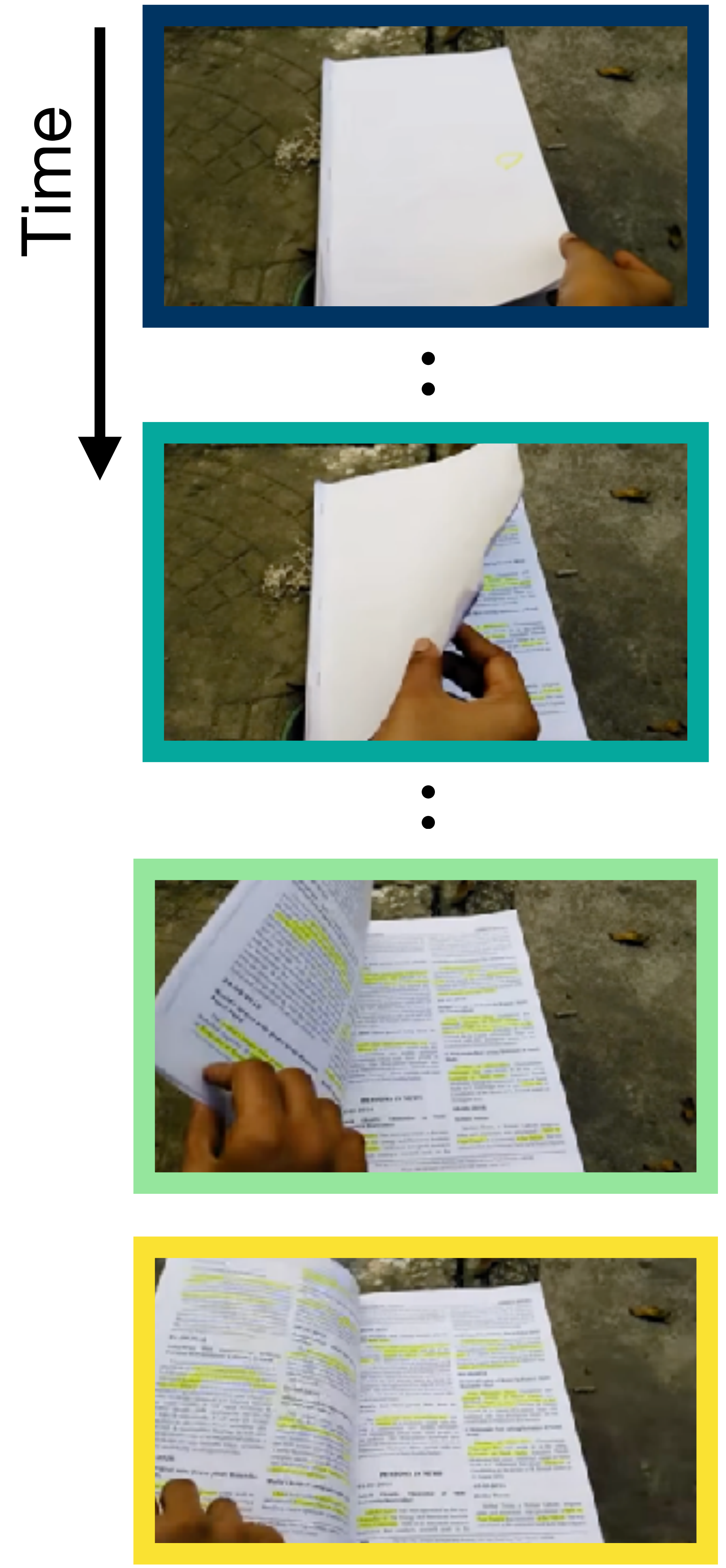
LiFT: A time-aware video embedding

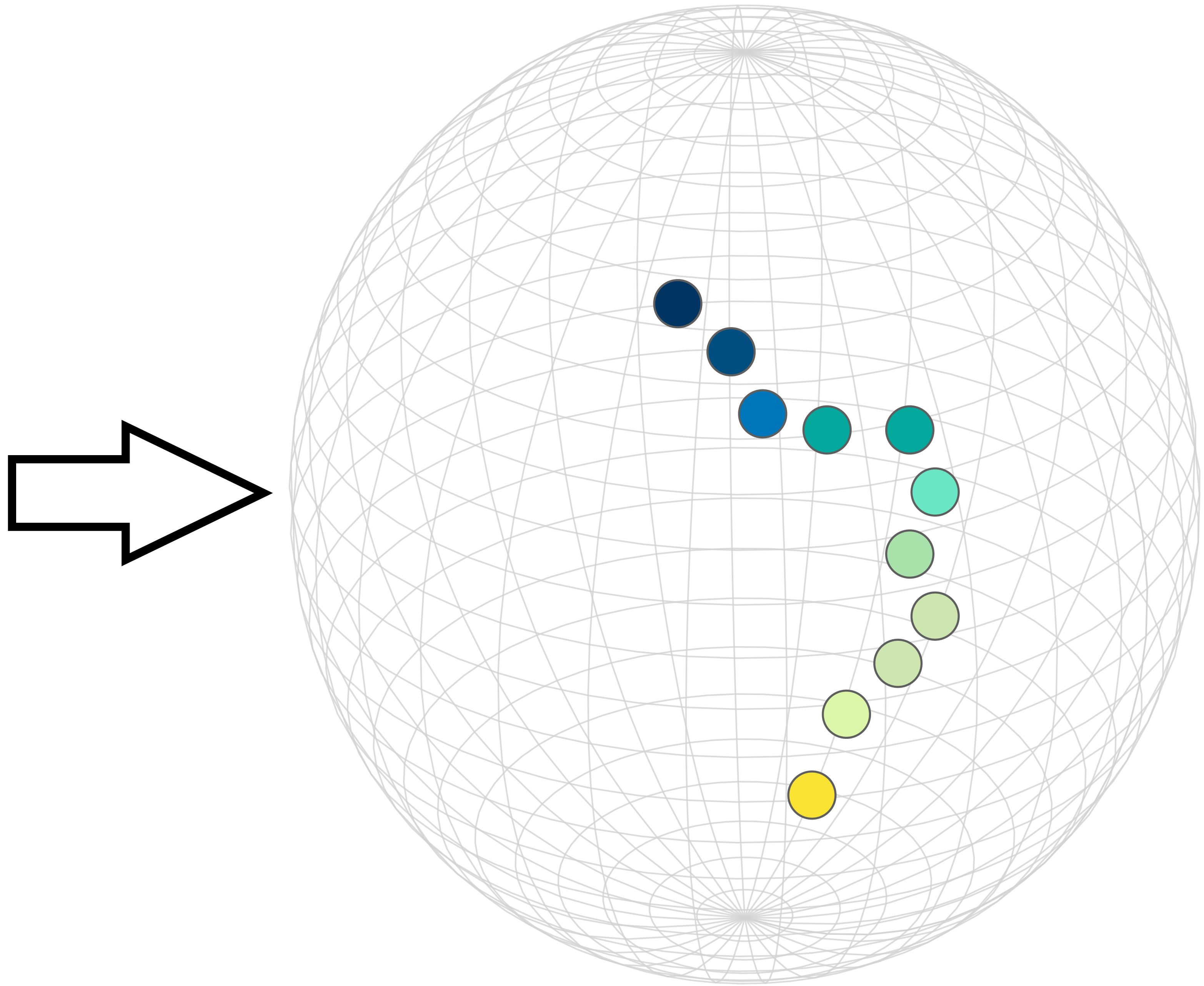
Encoding



LiFT: A time-aware video embedding

Encoding



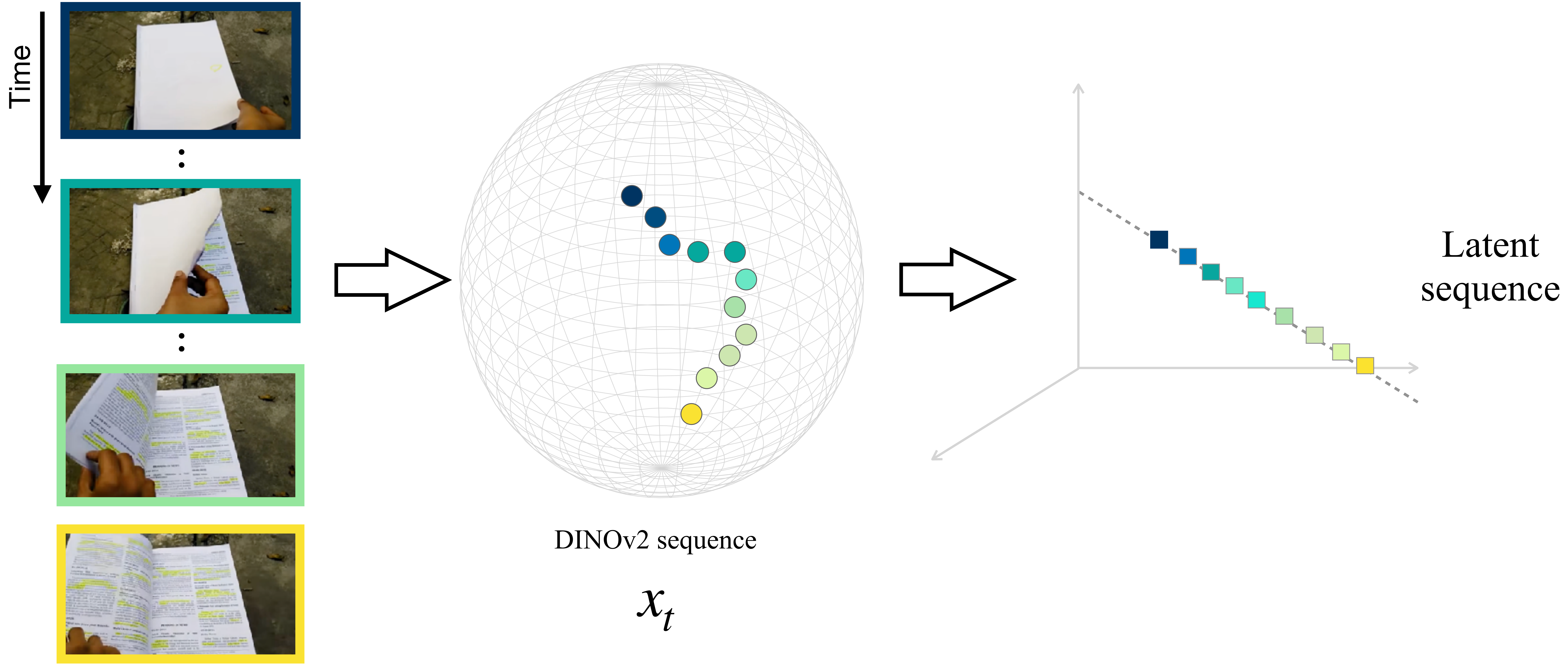


DINOv2 sequence

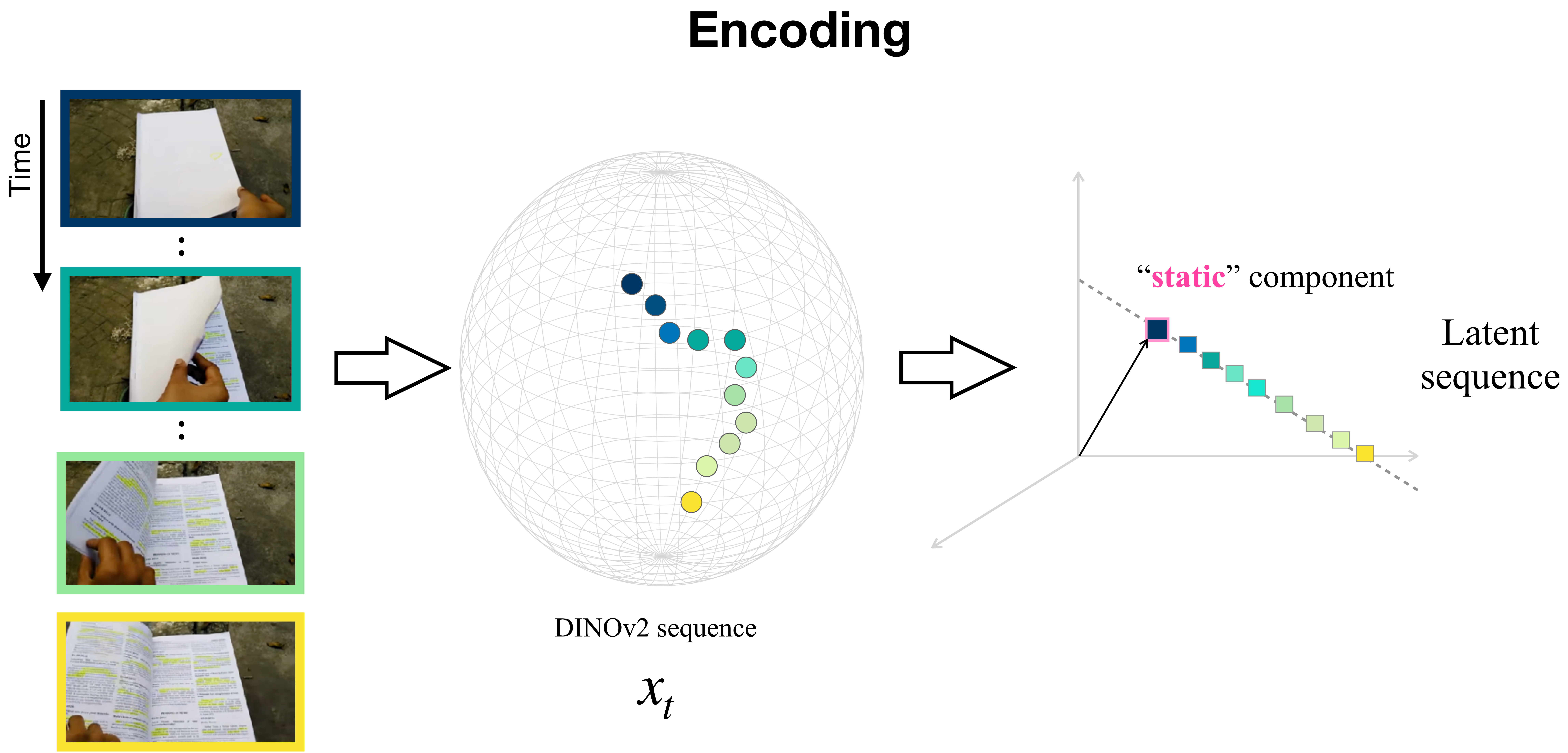
x_t

LiFT: A time-aware video embedding

Encoding

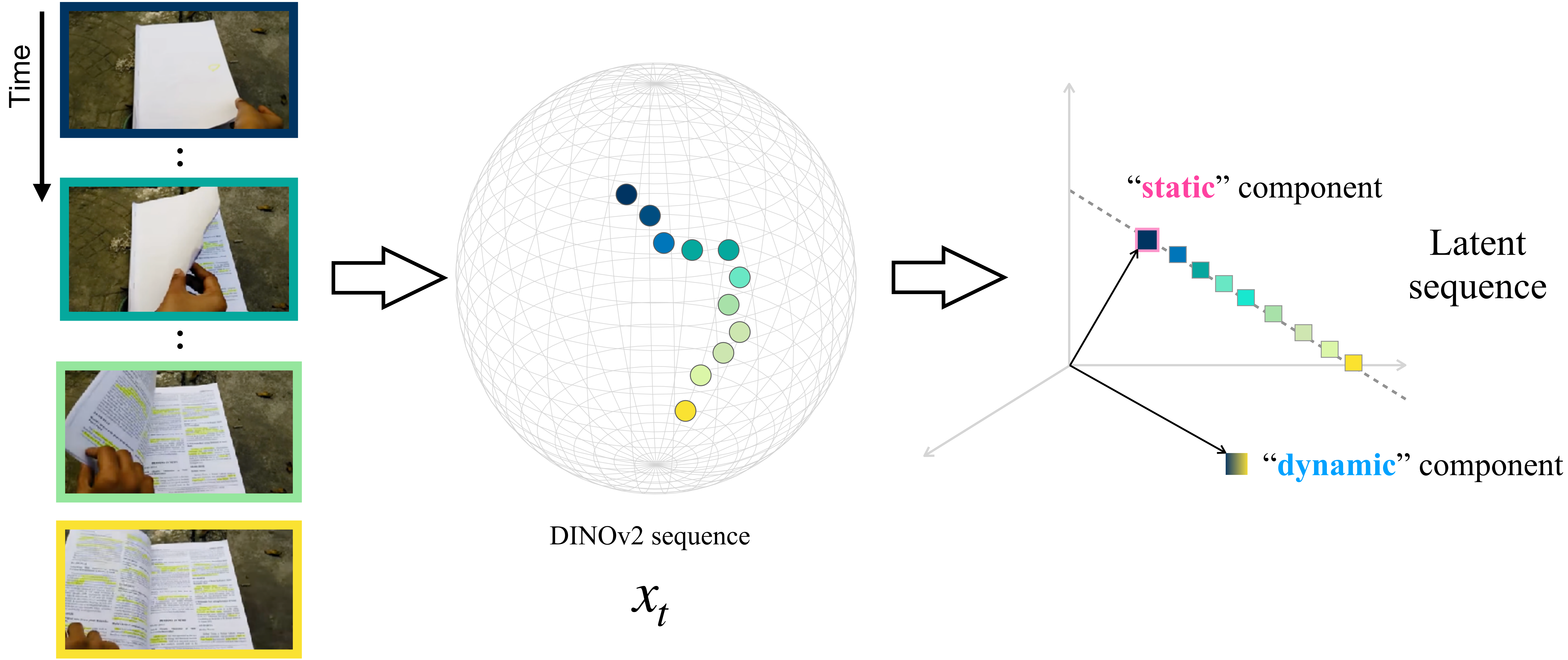


LiFT: A time-aware video embedding



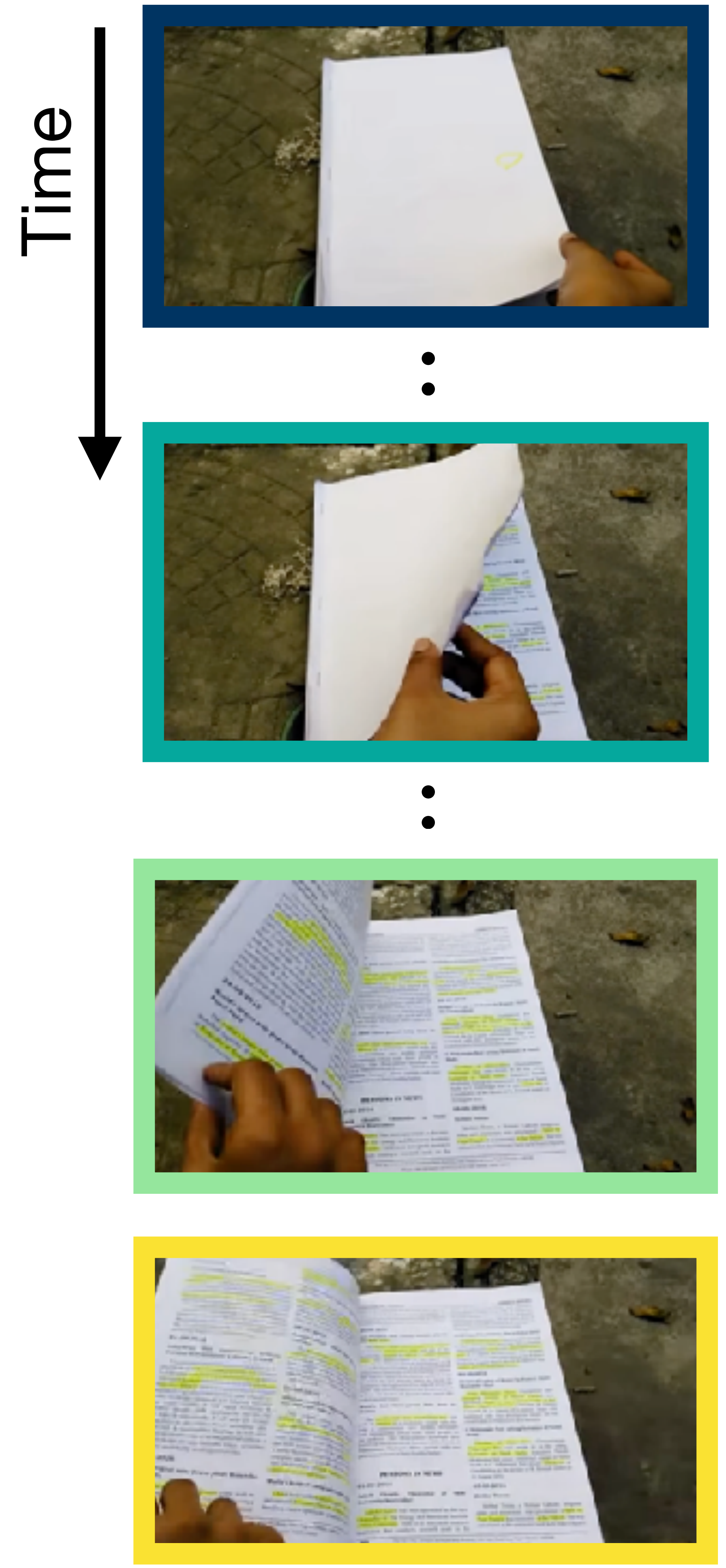
LiFT: A time-aware video embedding

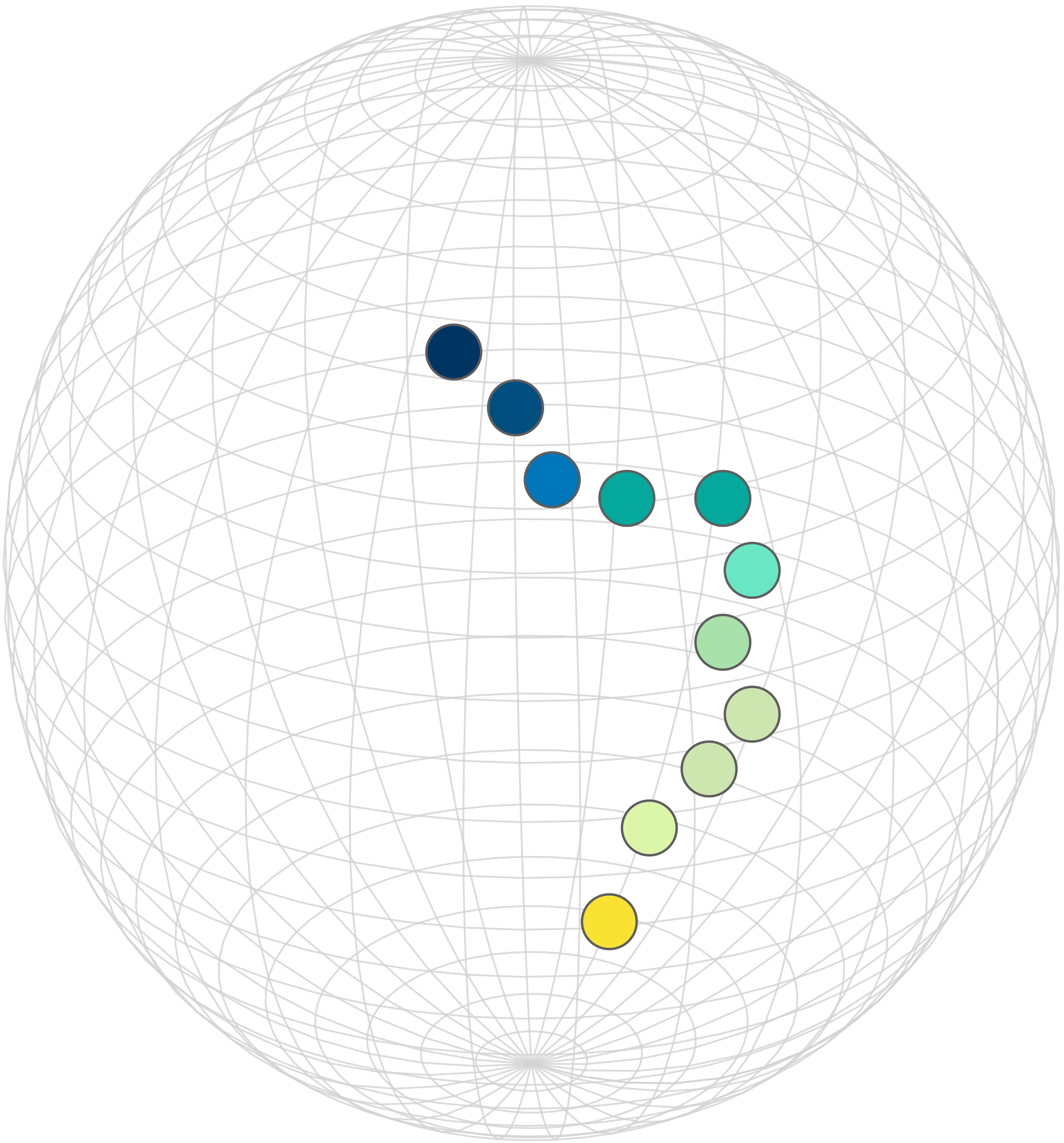
Encoding



LiFT: A time-aware video embedding

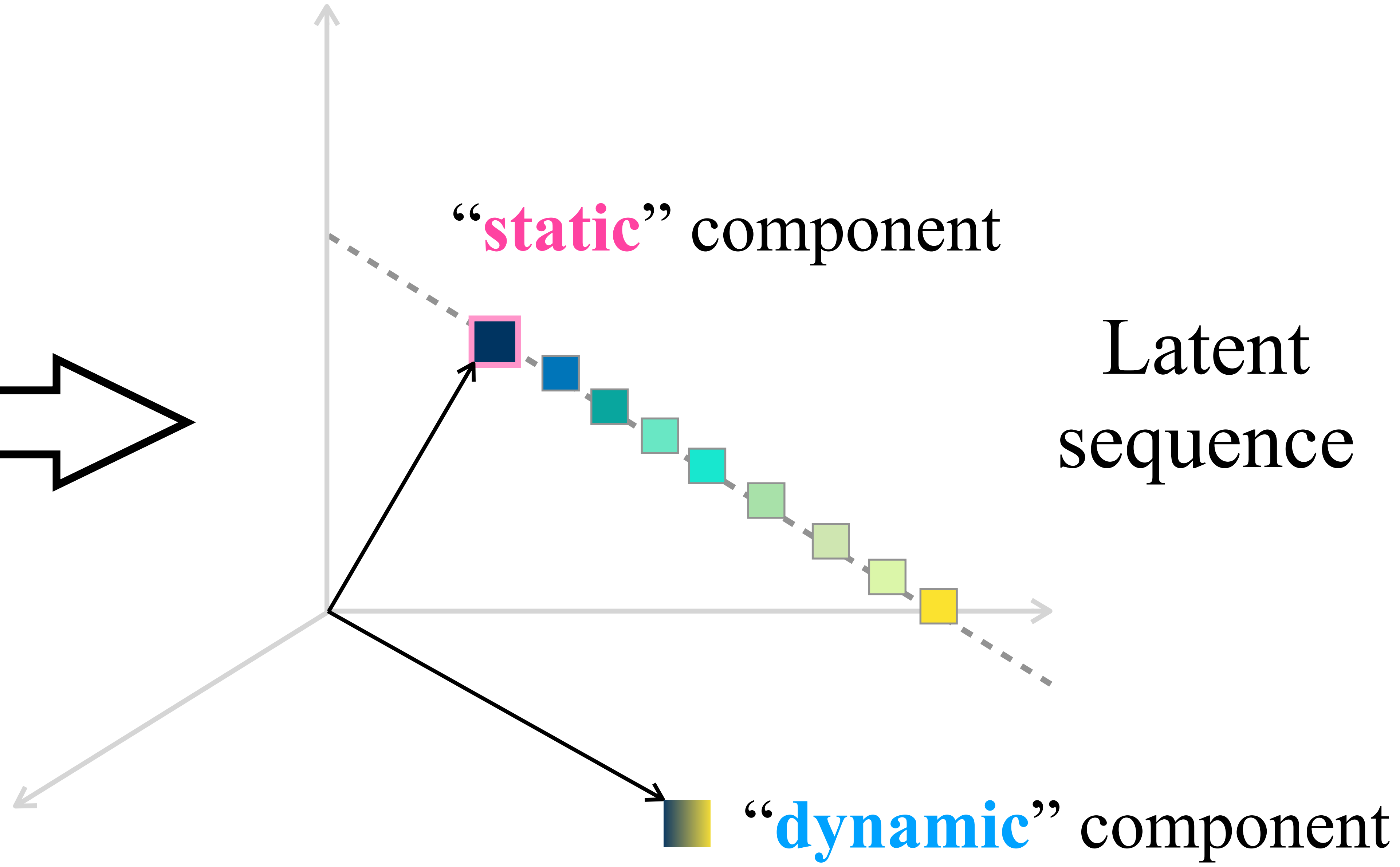
Encoding





DINOv2 sequence

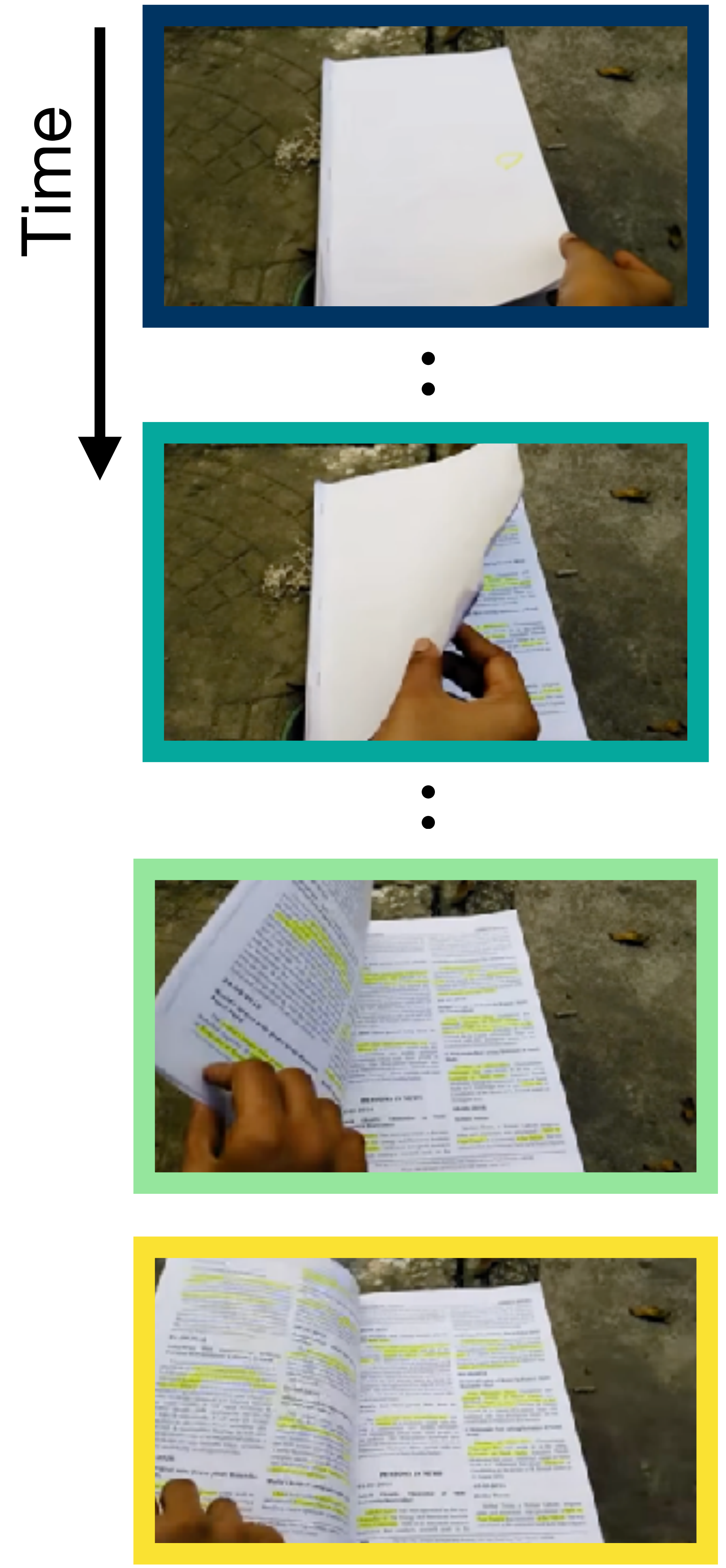
$$x_t$$

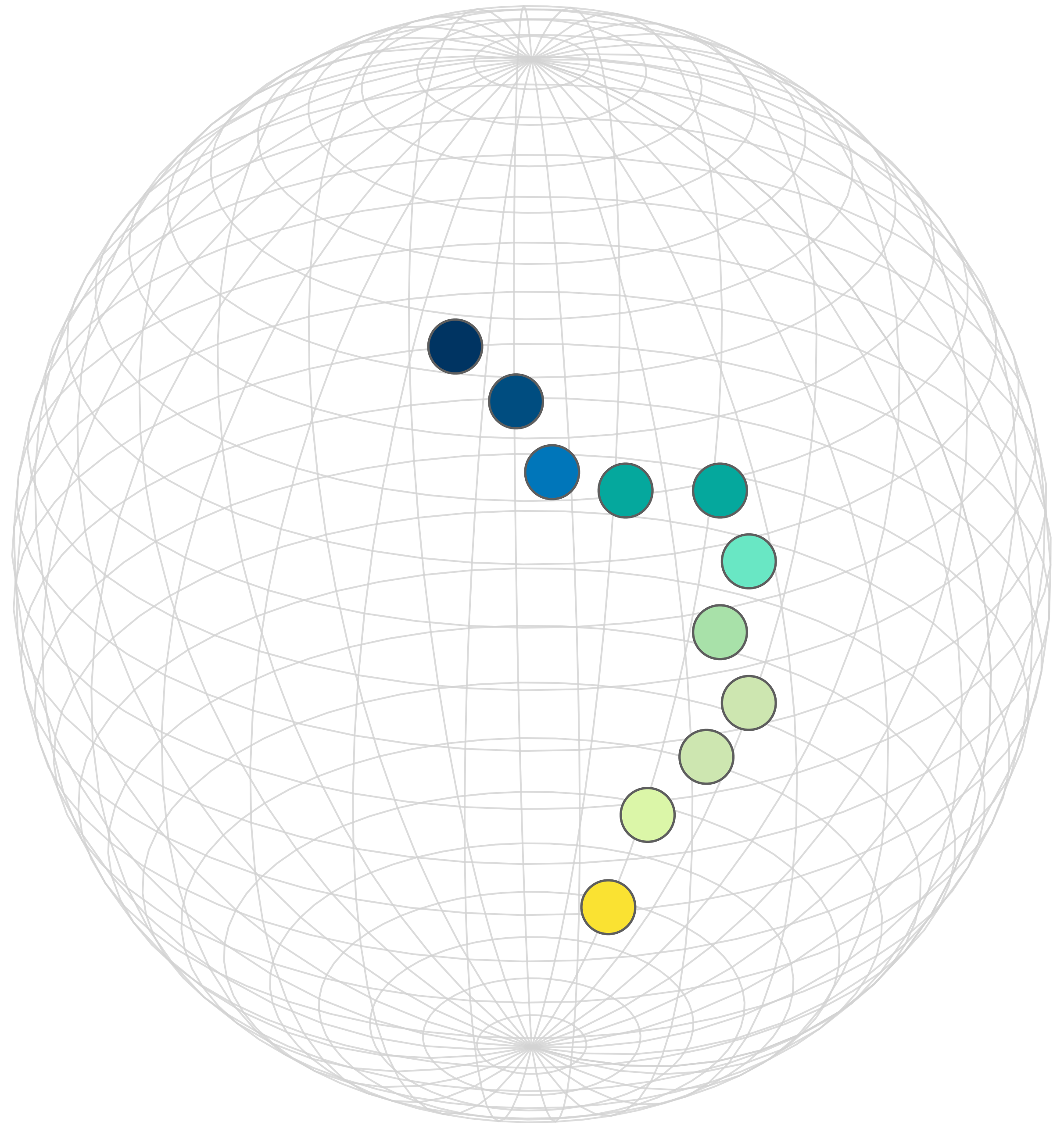


$$z_t := z_s + \left(\frac{t}{T}\right) z_d$$

LiFT: A time-aware video embedding

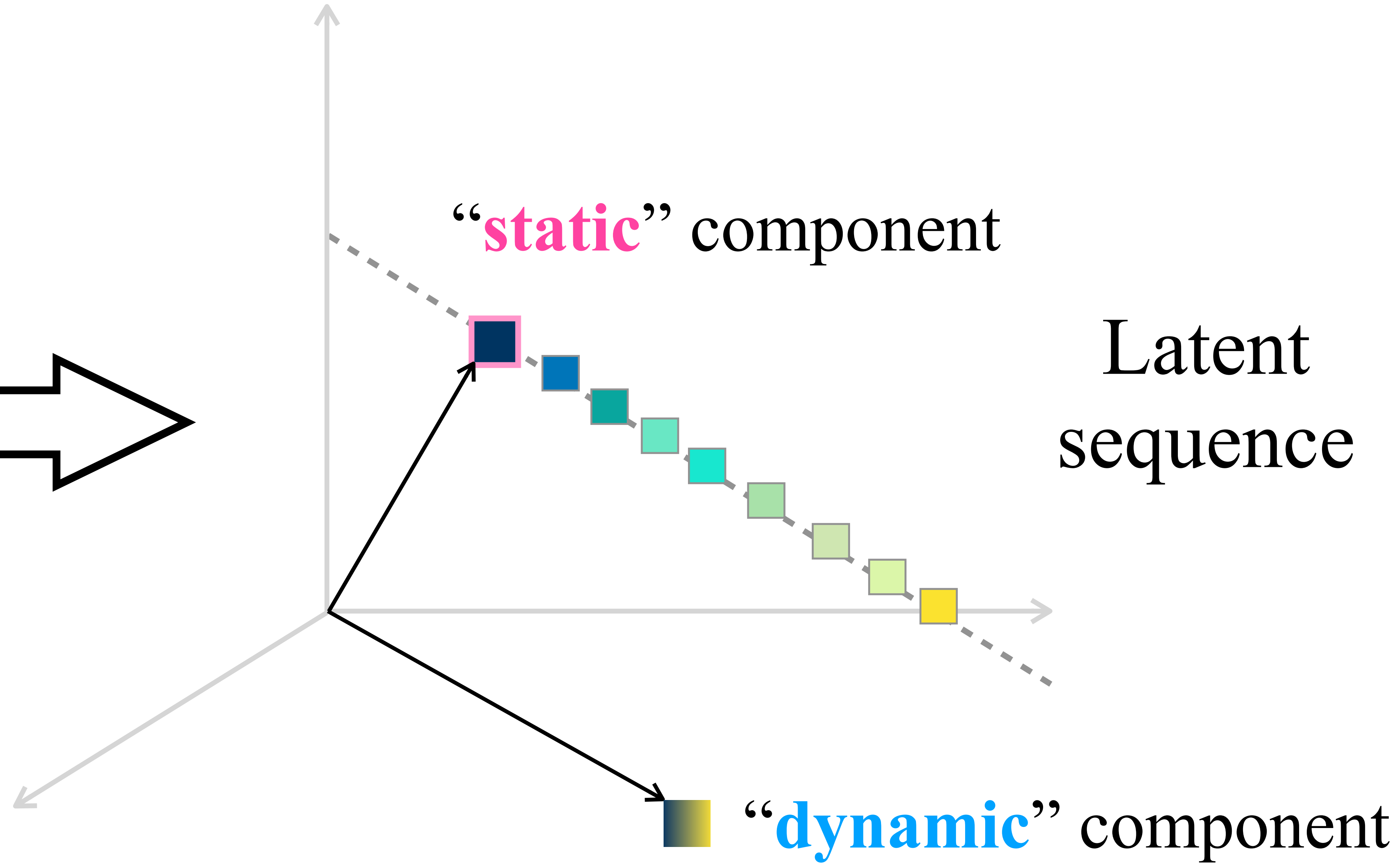
Encoding





DINOv2 sequence

$$x_t$$

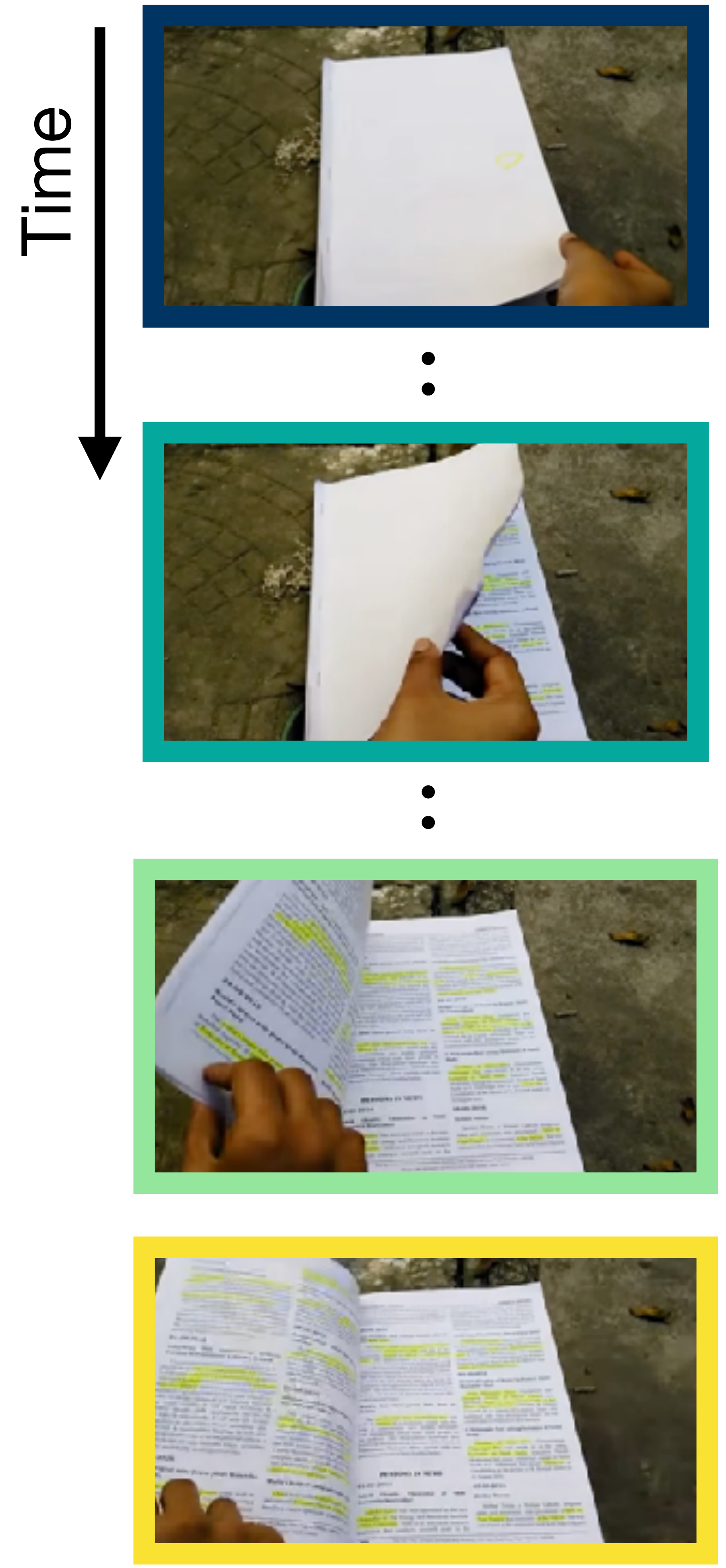


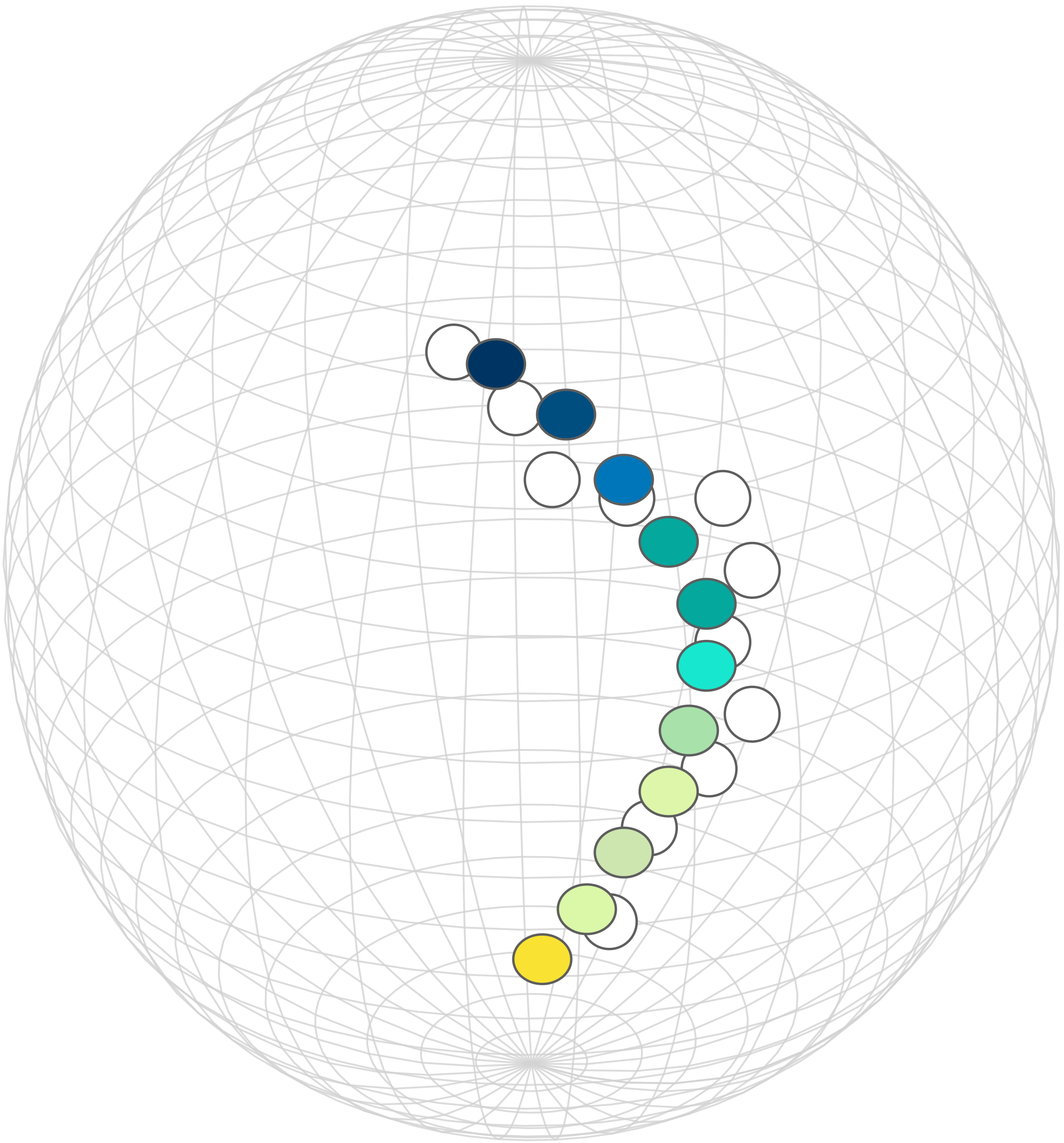
$$z_t := z_s + \left(\frac{t}{T}\right) z_d$$

Video sequence represented by (z_s, z_d)

LiFT: A time-aware video embedding

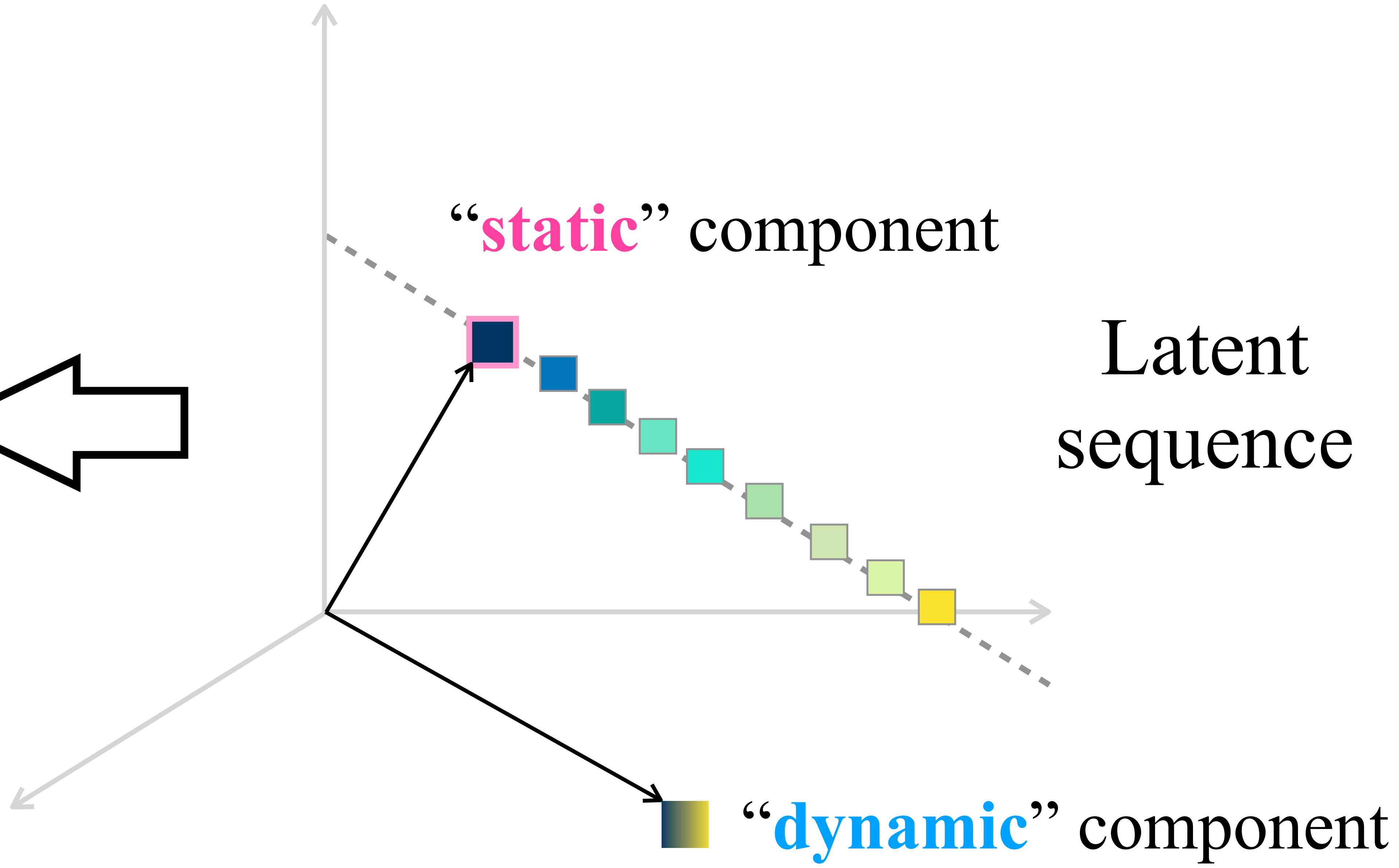
Decoding





Reconstructed DINOv2 sequence

$$\hat{x}_t$$

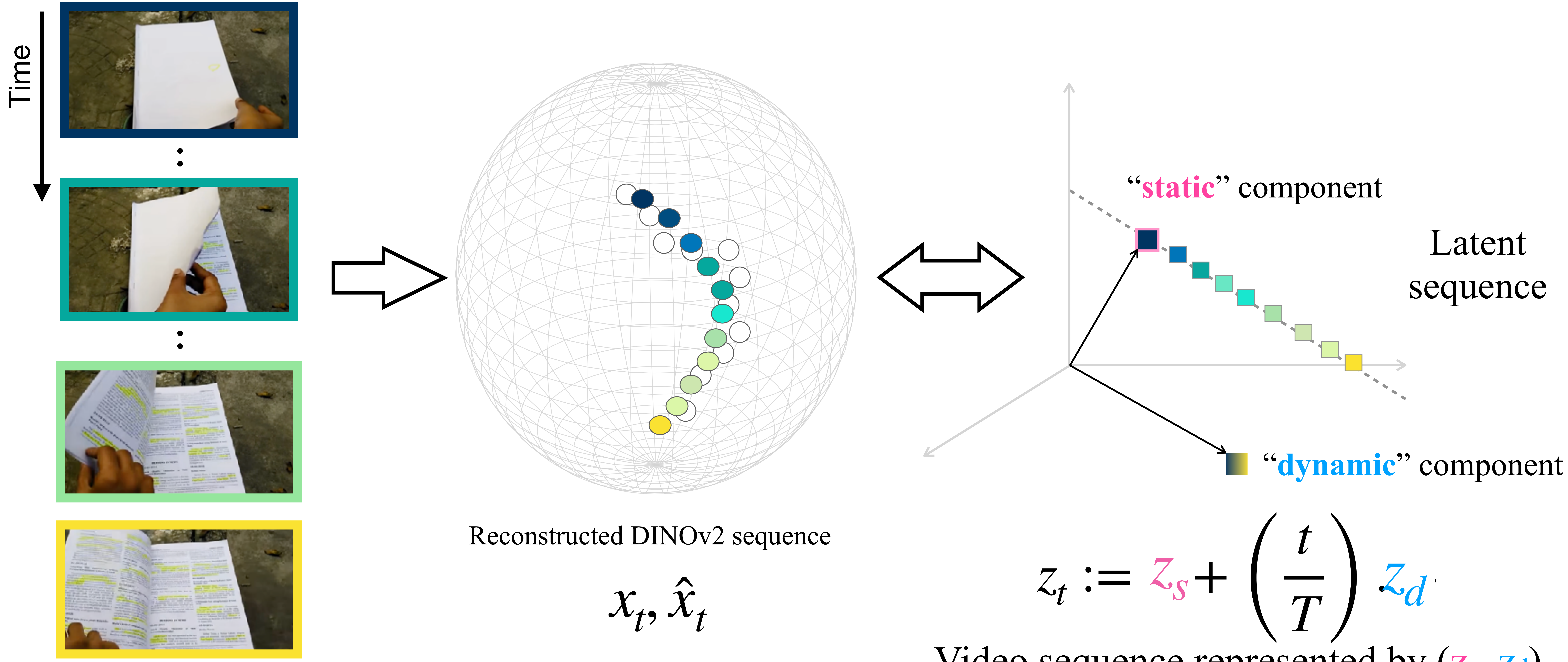


$$z_t := z_s + \left(\frac{t}{T}\right) z_d$$

Video sequence represented by (z_s, z_d)

LiFT: A time-aware video embedding

Auto-Encoder



Reconstructed DINOv2 sequence

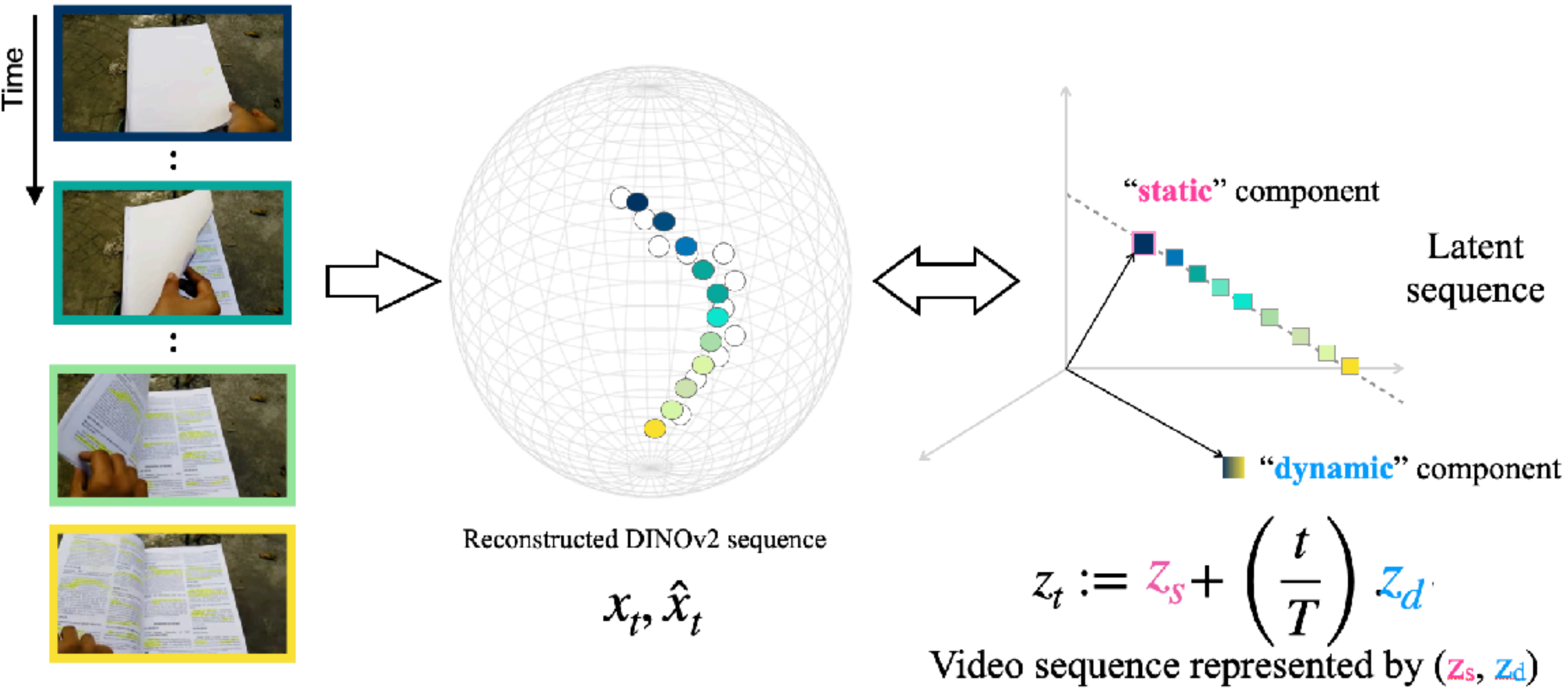
$x_t, \hat{x}_t$

$$z_t := z_s + \left(\frac{t}{T}\right) z_d$$

Video sequence represented by (z_s, z_d)

LiFT: A time-aware video embedding

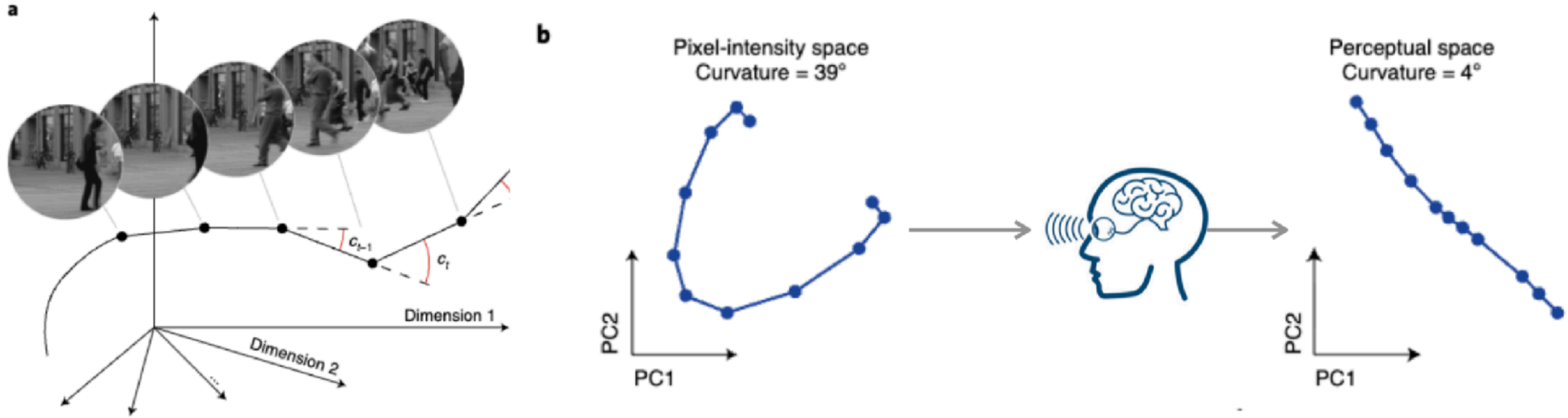
Auto-Encoder



- **Time-aware** by design: as it has to generate entire DINO sequence
- **Compact** as dimension of latents << dimension of temporal DINO sequence
- **Simple**: feature trajectory is mapped to a linearised space

LiFT is loosely inspired by “Perceptual Straightening Hypothesis”

Henaff et al. (2019) hypothesized that humans convert non-linear spatial representations of naturally occurring videos into linear temporal trajectories.



[1] Perceptual straightening of natural videos. Olivier J. Hénaff, Robbe L. T. Goris and Eero P. Simoncelli. Nature 2019.

LiFT: *Self-supervised* training

$$\mathcal{L} := \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{orth}} = \sum_{t=1}^{T} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2 + \lambda.\,\text{cos-sim}\left(\frac{\mathbf{z}_s}{\|\mathbf{z}_s\|_2}, \frac{\mathbf{z}_d}{\|\mathbf{z}_d\|_2}\right)$$

- Trained on 240K videos from Kinetics-400 with usual reconstruction loss and an orthogonality regularisation
- LiFT can be trained in < 1 day on a single GPU

Chirality in Action (CiA) Benchmark

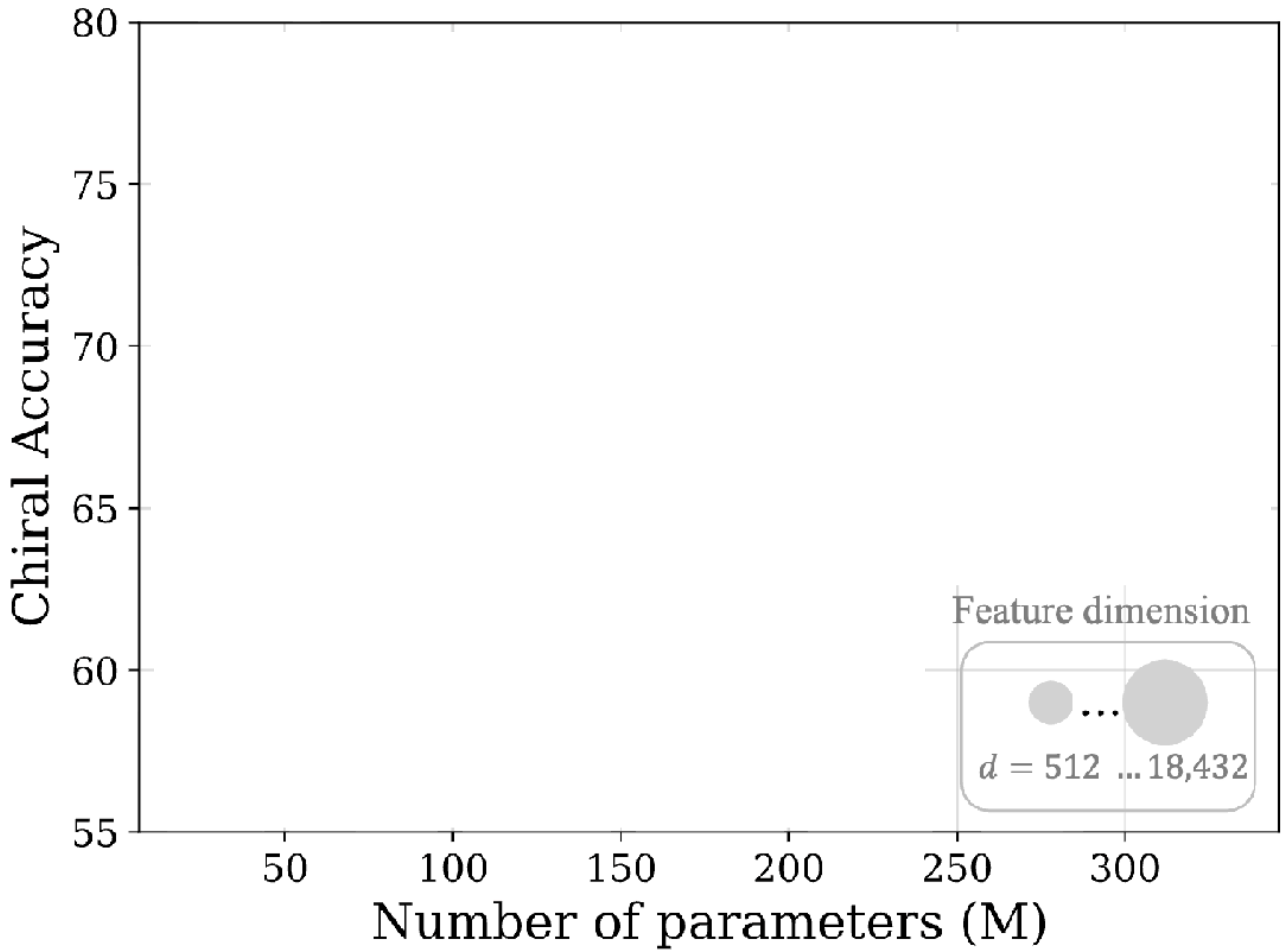
- (Meta) dataset to probe temporal ability of video embeddings
- Steps:
  1. Come up with temporal antonym verb pairs (e.g., open/close, move up/move down, etc.)
  2. Mine 3 datasets (**Something-something v2, EPIC, Charades**) for such pairs
  3. Manually review and filter

Base dataset	Chiral groups	Avg videos/group	Example chiral group
Something-Something (SSv2)	16	852.8	Folding / Unfolding [something]
EPIC-Kitchens (EPIC)	66	412.2	Opening / Closing [door]
Charades	28	768.4	Taking / Putting a [laptop]

Chirality in Action (CiA): Examples

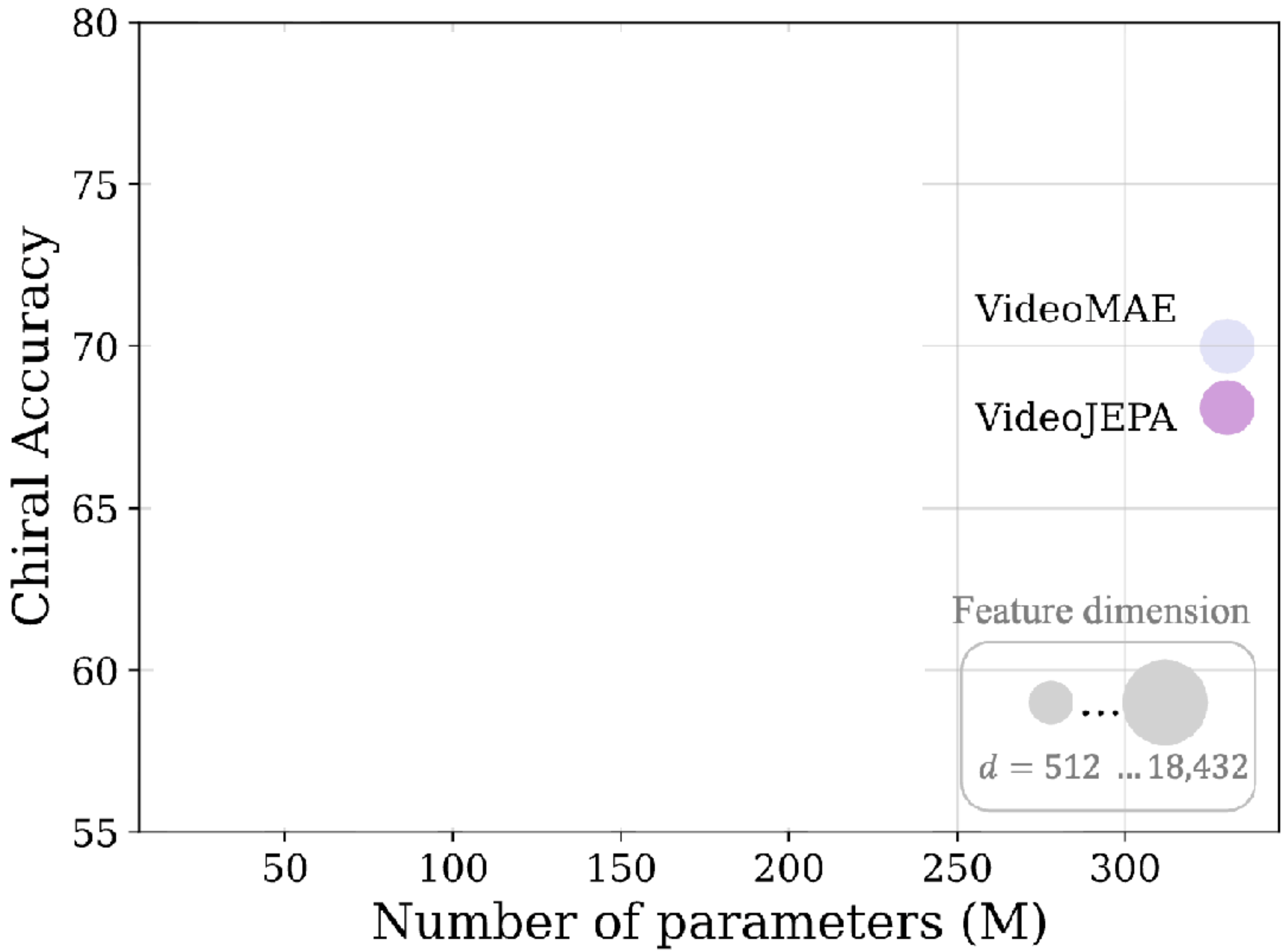
Evaluation on CiA (average across datasets)

- Chance is at 50%



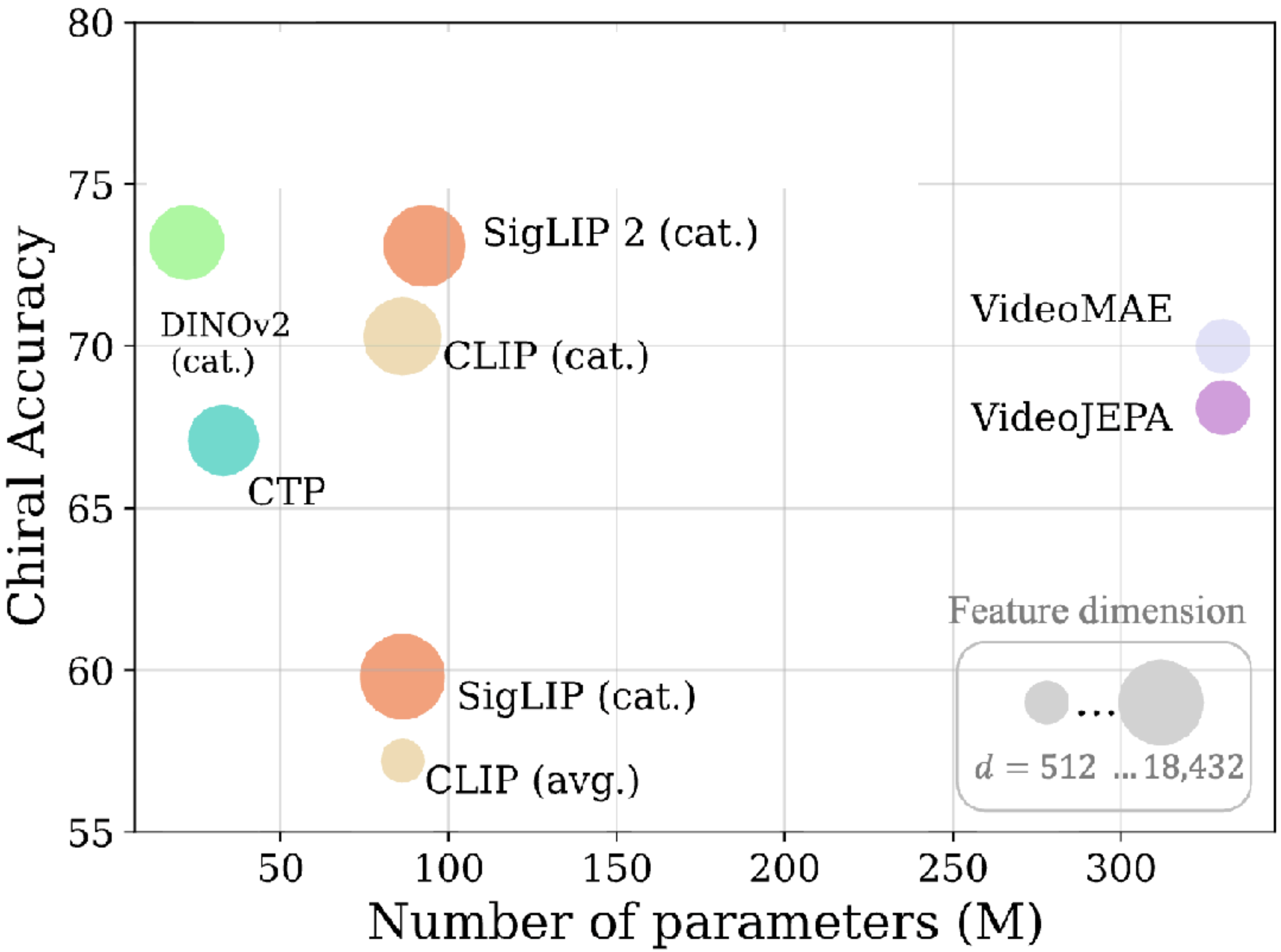
Evaluation on CiA (average across datasets)

- Chance is at 50%
- Video models like VideoJEPA perform well but are heavy



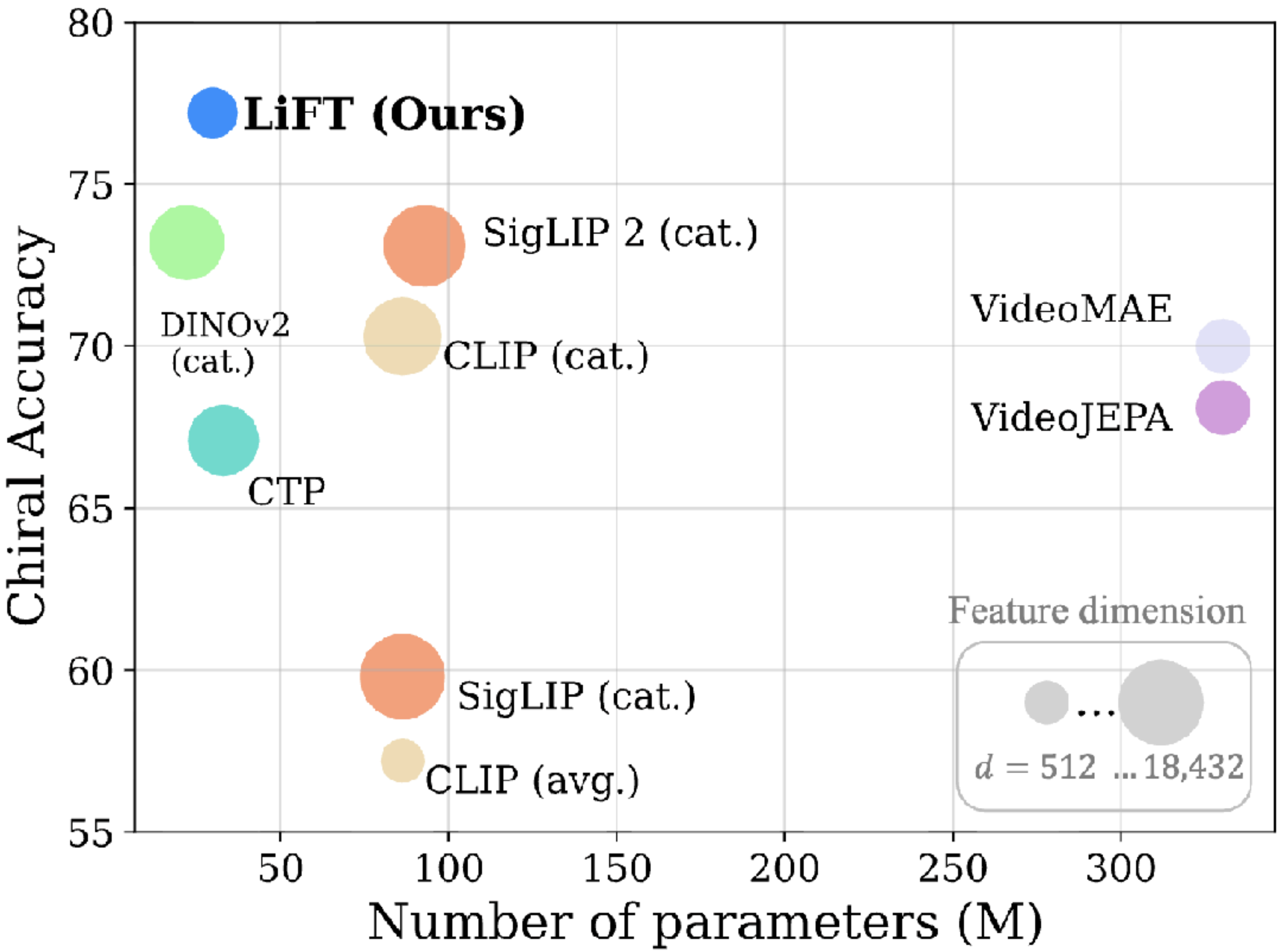
Evaluation on CiA (average across datasets)

- Chance is at 50%
- Video models like VideoJEPA perform well but are heavy
- Concatenating all image features (e.g., DINOv2) is very bulky



Evaluation on CiA (average across datasets)

- Chance is at 50%
- Video models like VideoJEPA perform well but are heavy
- Concatenating all image features (e.g., DINOv2) is very bulky
- LiFT outperforms all of them while being compact & has fewer parameters



Evaluation on standard benchmarks

- **Setup:** linear probe evaluation on standard benchmarks

Model	K400	UCF	HMDB	SSv2
Chance	0.25	0.99	1.96	0.58

Evaluation on standard benchmarks

- **Setup:** linear probe evaluation on standard benchmarks
- LiFT by itself achieves reasonable % but is not state-of-the-art

Model	K400	UCF	HMDB	SSv2
Chance	0.25	0.99	1.96	0.58
LiFT	55.4	86.6	65.2	30.8

Evaluation on standard benchmarks

- **Setup:** linear probe evaluation on standard benchmarks
- LiFT by itself achieves reasonable % but is not state-of-the-art
- However, LiFT when combined with existing models achieves state-of-the-art performance

Model	**K400**	**UCF**	**HMDB**	**SSv2**
Chance	0.25	0.99	1.96	0.58
LiFT	55.4	86.6	65.2	30.8
VJEPA	59.8†	91.3	76.1	49.6†
VJEPA $\oplus$ LiFT	63.7	92.6	78.0	52.3
Δ	+3.9	+1.3	+1.9	+2.7

Evaluation on standard benchmarks

- **Setup:** linear probe evaluation on standard benchmarks
- LiFT by itself achieves reasonable % but is not state-of-the-art
- However, LiFT when combined with existing models achieves state-of-the-art performance

Model	K400	UCF	HMDB	SSv2
Chance	0.25	0.99	1.96	0.58
LiFT	55.4	86.6	65.2	30.8
VJEPA	$59.8^{\dagger}$	91.3	76.1	$49.6^{\dagger}$
VJEPA $\oplus$ LiFT	63.7	92.6	78.0	52.3
Δ	+3.9	+1.3	+1.9	+2.7
VideoMAE	55.0	83.6	66.5	38.3
VideoMAE $\oplus$ LiFT	63.6	88.8	72.6	46.3
Δ	+8.6	+5.2	+6.1	+6.0

Evaluation on standard benchmarks

- **Setup:** linear probe evaluation on standard benchmarks
- LiFT by itself achieves reasonable % but is not state-of-the-art
- However, LiFT when combined with existing models achieves state-of-the-art performance

Model	K400	UCF	HMDB	SSv2
Chance	0.25	0.99	1.96	0.58
LiFT	55.4	86.6	65.2	30.8
VJEPA	59.8†	91.3	76.1	49.6†
VJEPA $\oplus$ LiFT	63.7	92.6	78.0	52.3
Δ	+3.9	+1.3	+1.9	+2.7
VideoMAE	55.0	83.6	66.5	38.3
VideoMAE $\oplus$ LiFT	63.6	88.8	72.6	46.3
Δ	+8.6	+5.2	+6.1	+6.0
InternVid2.5	62.8	88.2	71.9	23.4
InternVid2.5 $\oplus$ LiFT	65.9	90.3	75.3	35.9
Δ	+3.1	+2.1	+3.4	+11.5

Summary

- LiFT, a simple video embedding model:
    - Time-aware
    - Compact
    - Self-supervised
- CiA: a benchmark of chiral (temporally opposite) action pairs to probe video embedding models
- LiFT achieves strong performance on CiA but also lifts up performance of de-facto video encoders on standard benchmarks

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF OXFORD

Thank you!

Project page

https://bpiyush.github.io/lift-website/