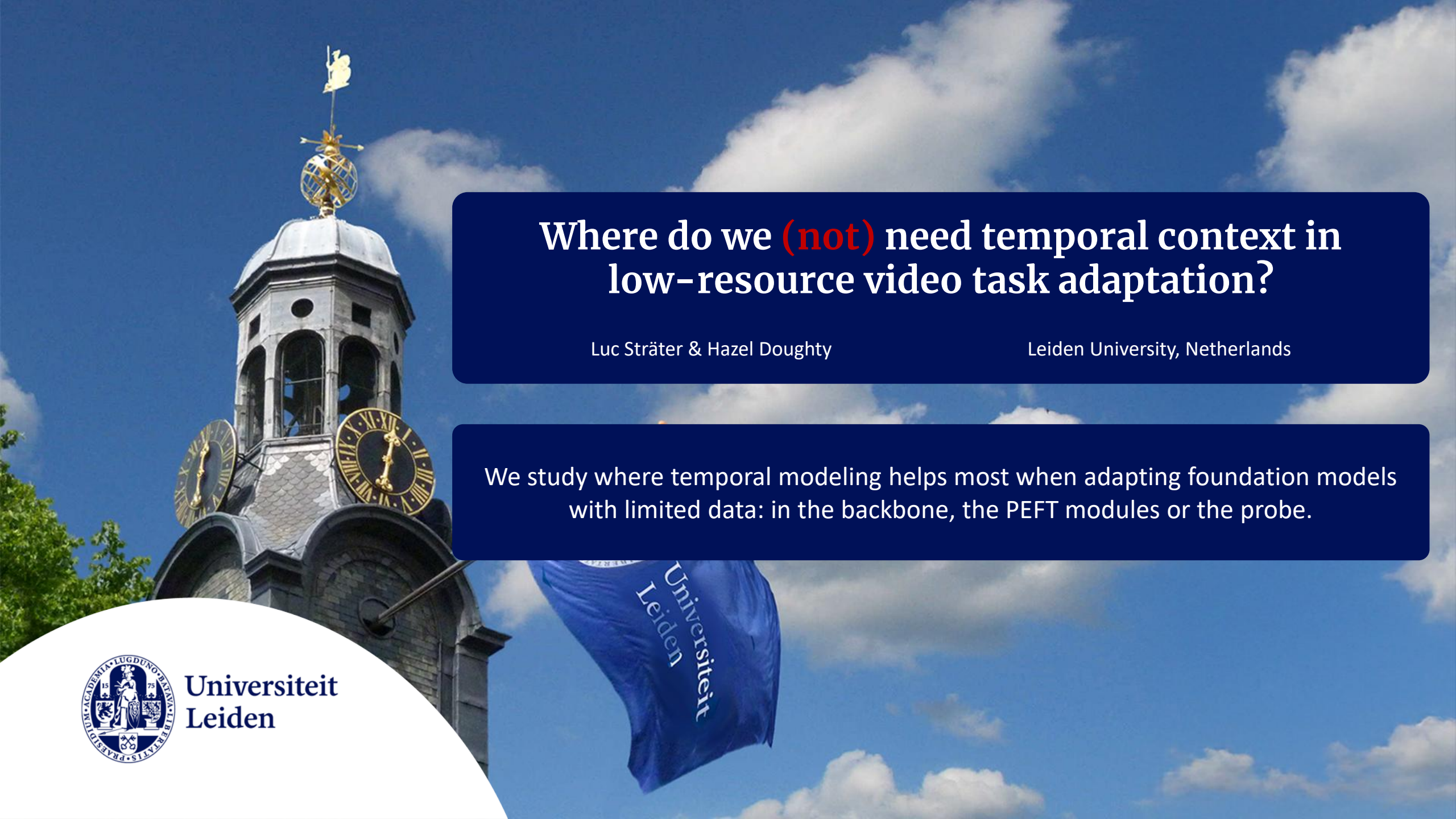




Where do we **(not)** need temporal context in low-resource video task adaptation?

Luc Sträter & Hazel Doughty

Leiden University, Netherlands

We study where temporal modeling helps most when adapting foundation models with limited data: in the backbone, the PEFT modules or the probe.

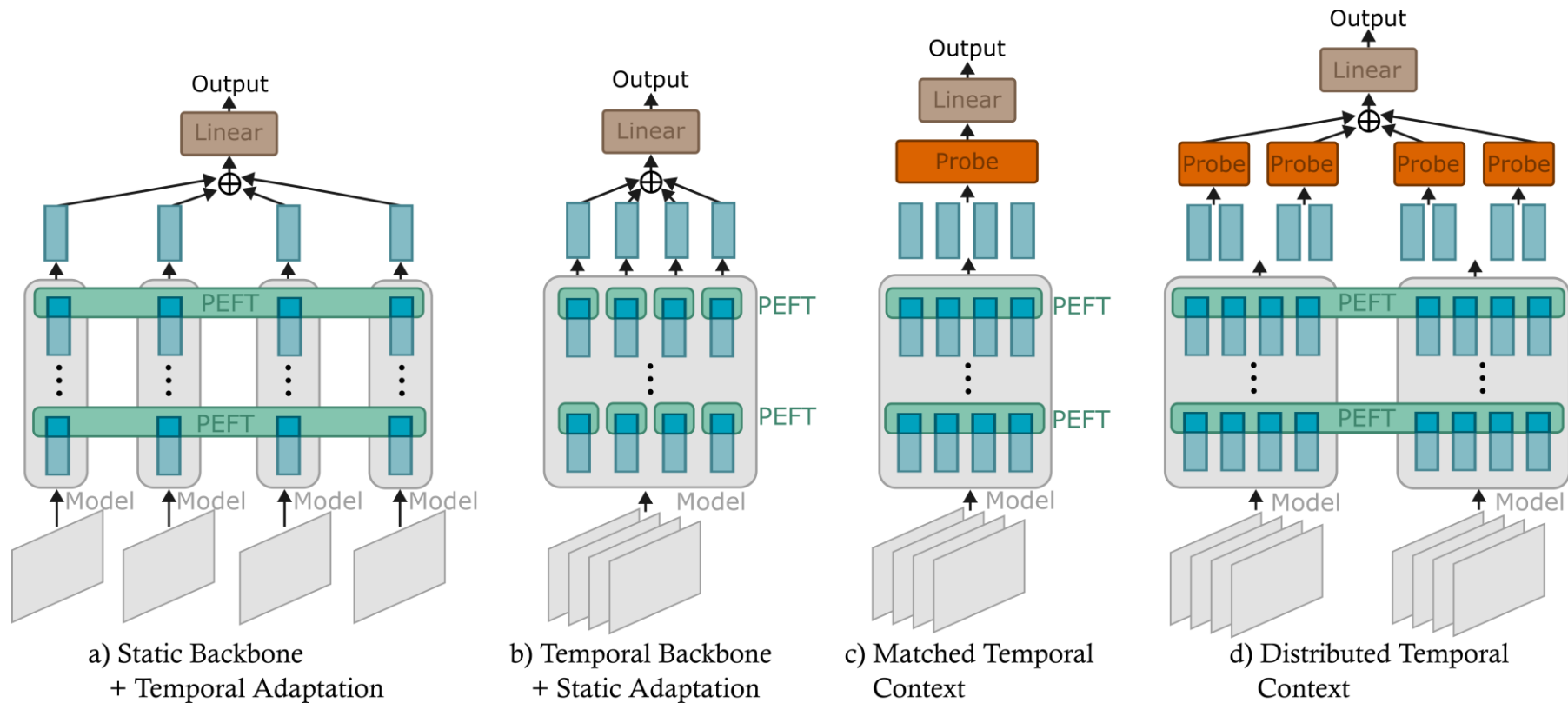


Universiteit
Leiden



Where should temporal reasoning happen?

Prior work models time in one component: the PEFT (a) or the backbone (b). **We model it in backbone, PEFT and probe; *matched*: all sharing one temporal context (c), or *distributed*: each with its own (d).**



Standard PEFT beats probing, at almost no cost

6 / 6

tasks where AdaptFormer + attentive probe beats the attentive probe alone

3%

throughput cost vs. probing alone (255.9 vs. 264.9 clips/s)

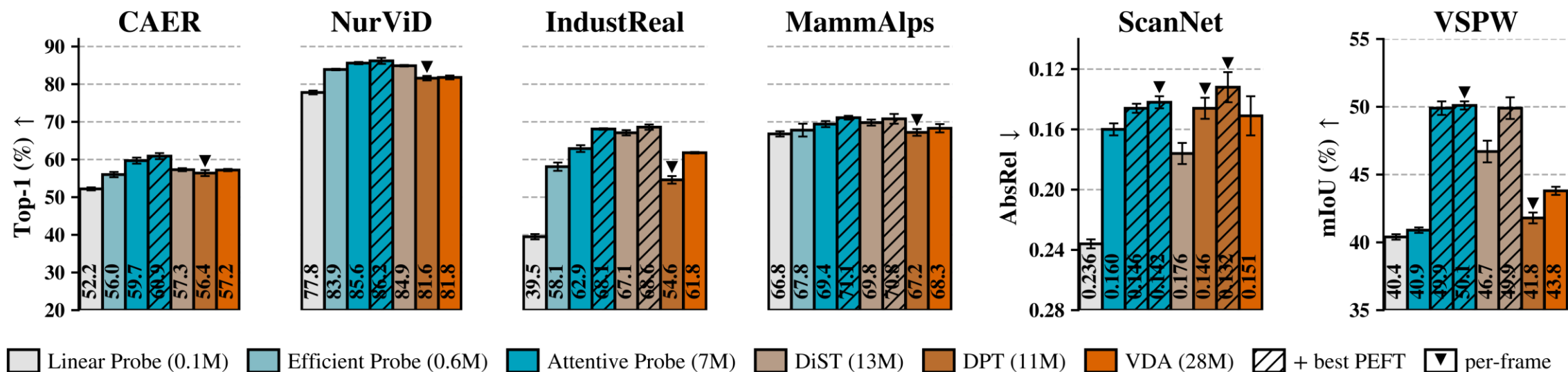
8×

fewer trainable parameters than full finetuning, at matching accuracy

- **Image-to-video PEFT does not pay off:** ST-Adapter adds parameters and slows inference, but never beats AdaptFormer beyond seed variance.
- **Video backbones win on motion:** for appearance and spatial tasks, the pretraining modality does not predict performance.

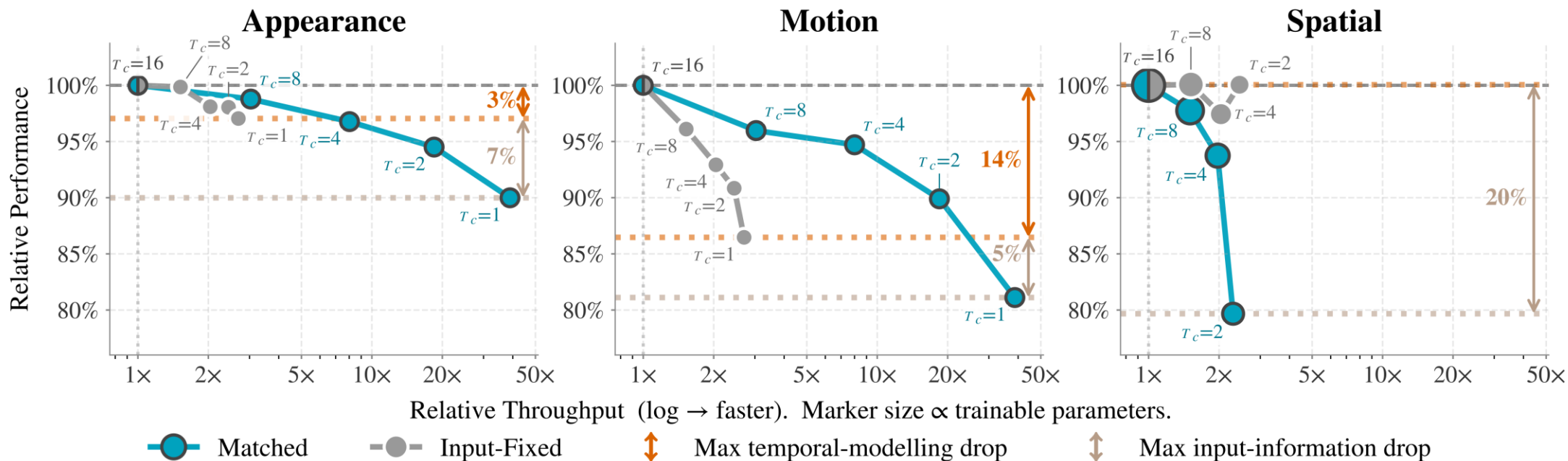
Which probe works best for video?

- Alone, the best probe is task-dependent
- With PEFT, a simple attentive probe suffices



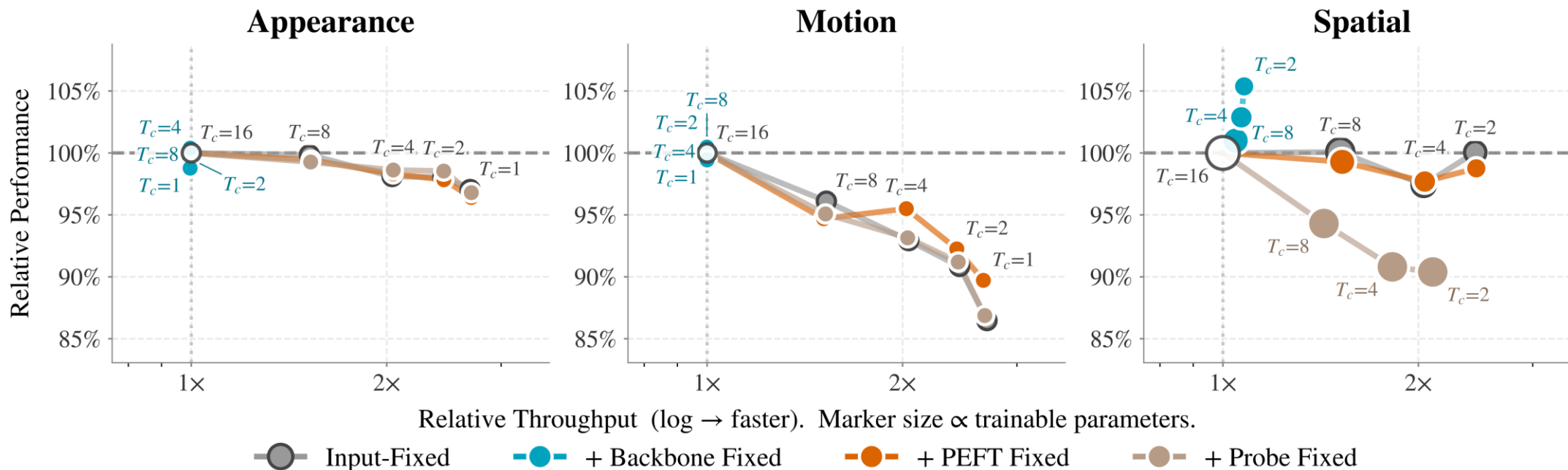
More frames, or temporal modeling?

- Motion relies on temporal modeling
- Appearance & spatial gain most from more frames



Where should the temporal context go?

- Temporal context matters most in the backbone
- For spatial tasks, temporal modeling in the probe hurts performance



Thank you!

Paper & results

lucstrater.com/temporal-context



Universiteit
Leiden

