Why Cross-Skeleton Retargeting Is Non-Identifiable

Structural Limits of Generative Motion Models

Zhiyuan Li[1] Wenyan Yang[1] Pekka Marttinen[2] Joni Pajarinen[1]

[1]Department of Electrical Engineering and Automation [2]Department of Computer Science

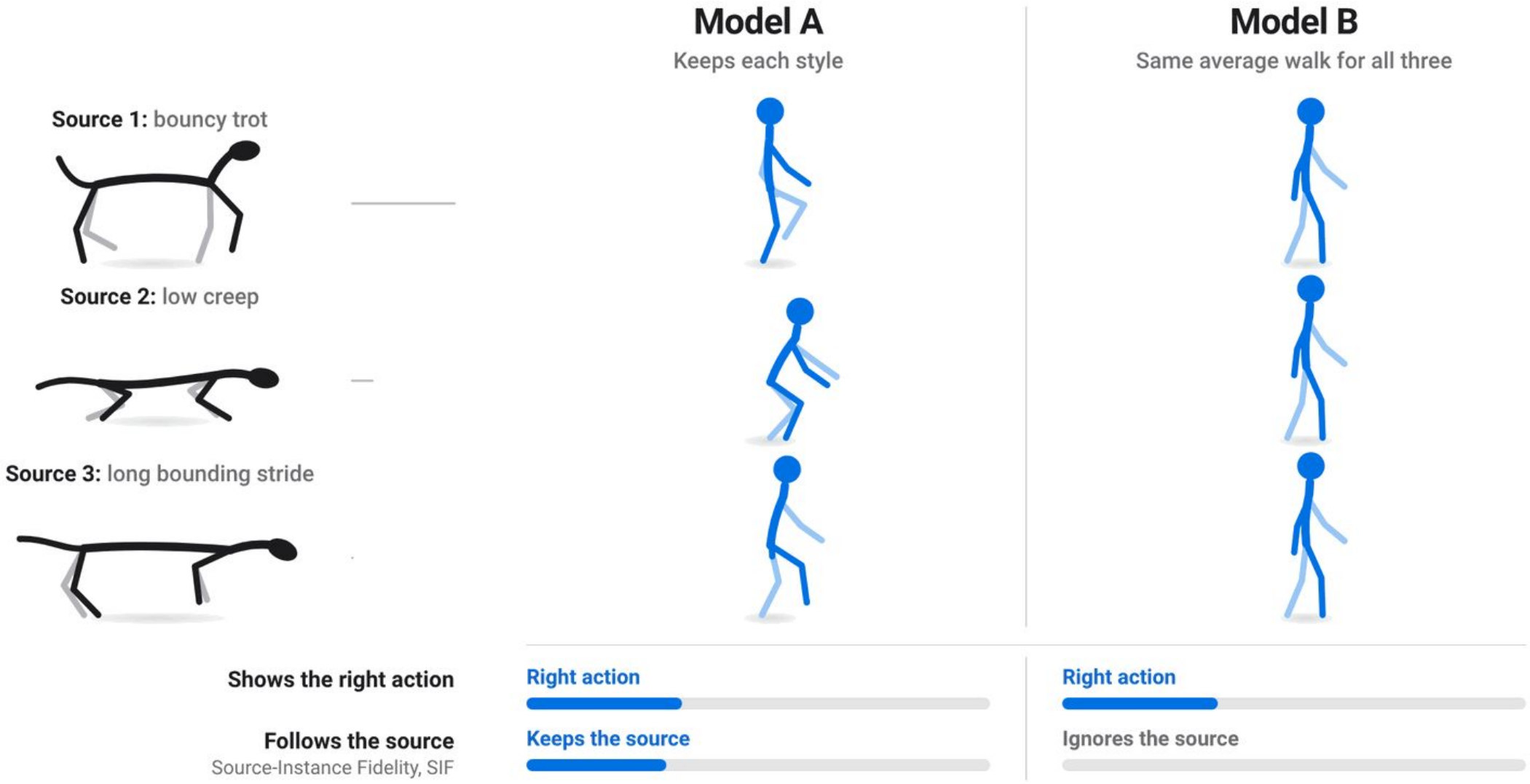
Aalto University, Finland

Aalto University

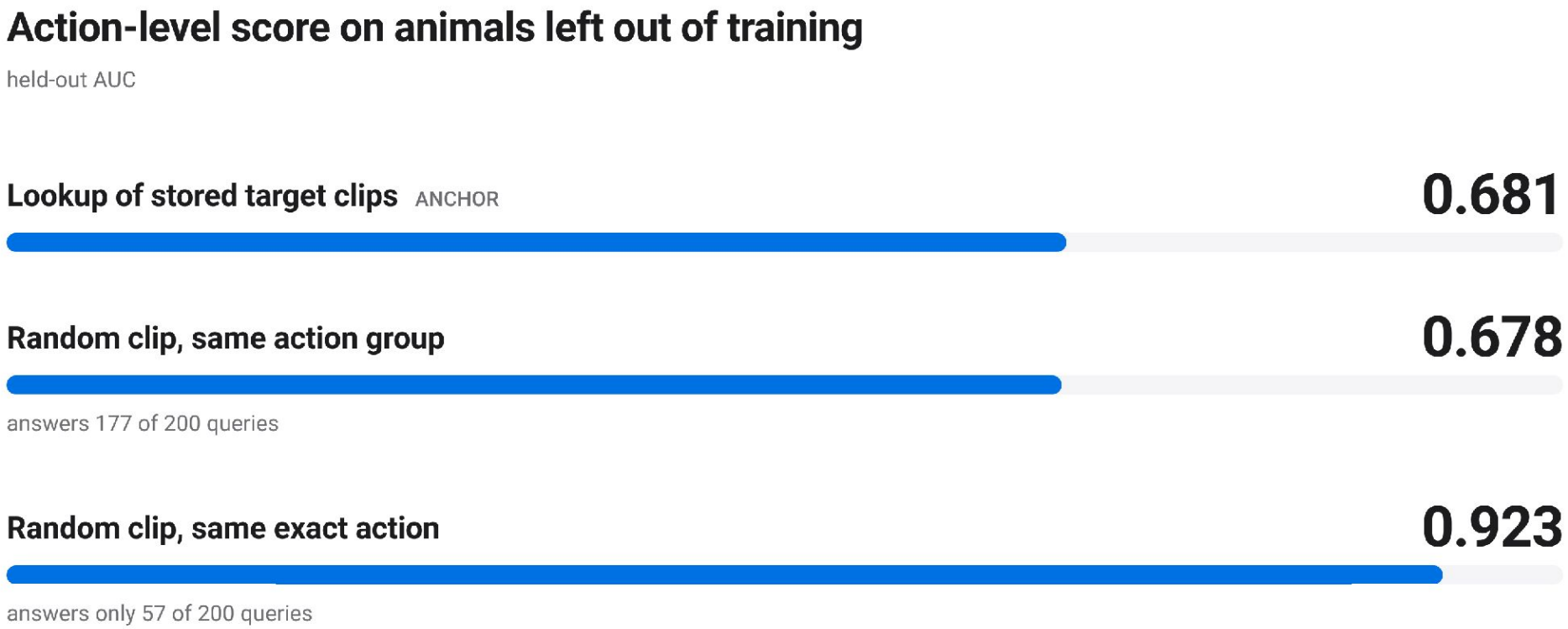
FCAI Finnish Center for Artificial Intelligence

NEURAL INFORMATION PROCESSING SYSTEMS

Cross-Skeleton Retargeting



Action-Level Evaluation



A random clip with the right label scores as well.

Truebones benchmark (Gat et al., 2025): retrieval AUC of the output against real target-body clips, same action = positive. Held-out skeletons, paper Table 1.

Gauge Non-Identifiability

Each dot is one motion. Training sees only how each body's motions are spread out.

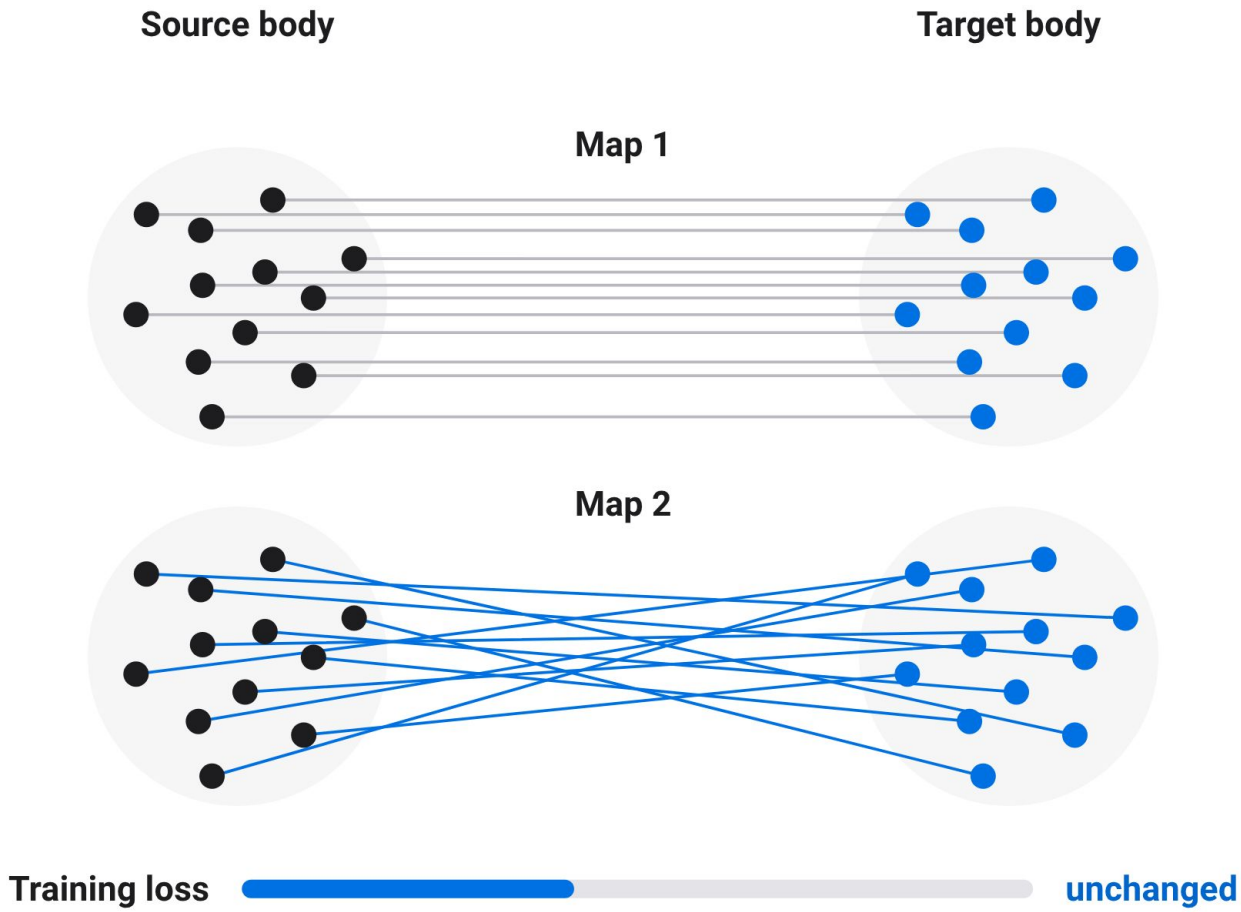
Turn the target body's dots. Their shape does not change.

The training loss does not change either.

But every source dot now lands on a different target dot.

$$\mathcal{L}(E, D) = \sum_{s=1}^{K} \mathcal{L}_s\big(D_s \circ E_s;\, P_s\big) + \Omega\big(\mu_1, \ldots, \mu_K\big)$$

Theorem 1. The objective fixes the retargeting map only up to a relative transformation of the bodies' latent spaces, such as a rotation.



Conditional-Mean Degeneration

One source clip. Several target clips of the same action would all be valid.

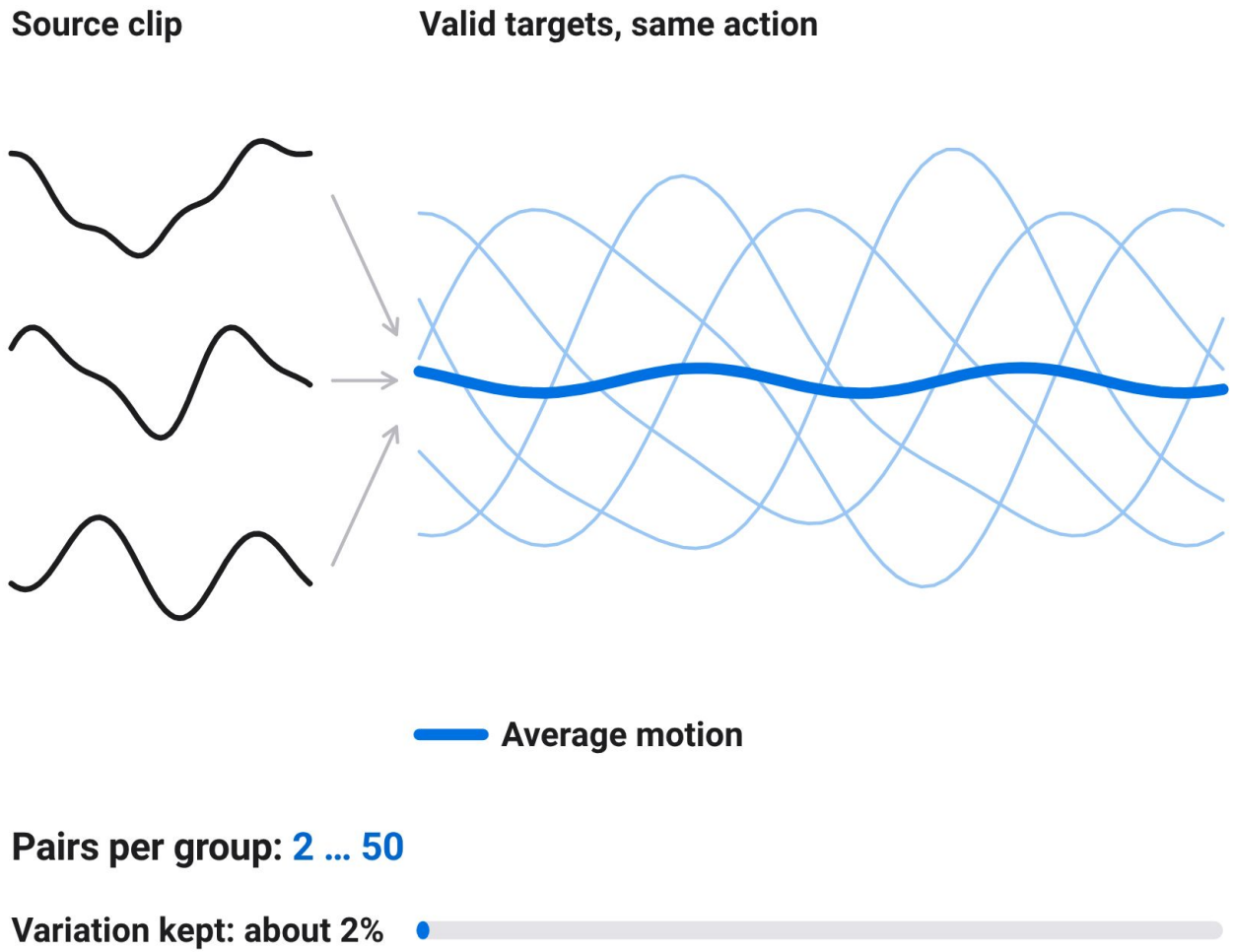
Each training step pairs it with one of them at random.

The model's answer is pulled toward the middle.

It settles on the average: the same answer for every source clip.

$$\mathcal{L}_{\text{paired}}(f) \;=\; \mathbb{E}\,\big\|\, f(x_a, b) \;-\; x_b \,\big\|_2^2$$

Proposition 1.1. Under random same-action pairing, the squared-error minimiser is the cell mean, independent of the source.



Source-Instance Fidelity (SIF)

Three source clips: same body, same action.

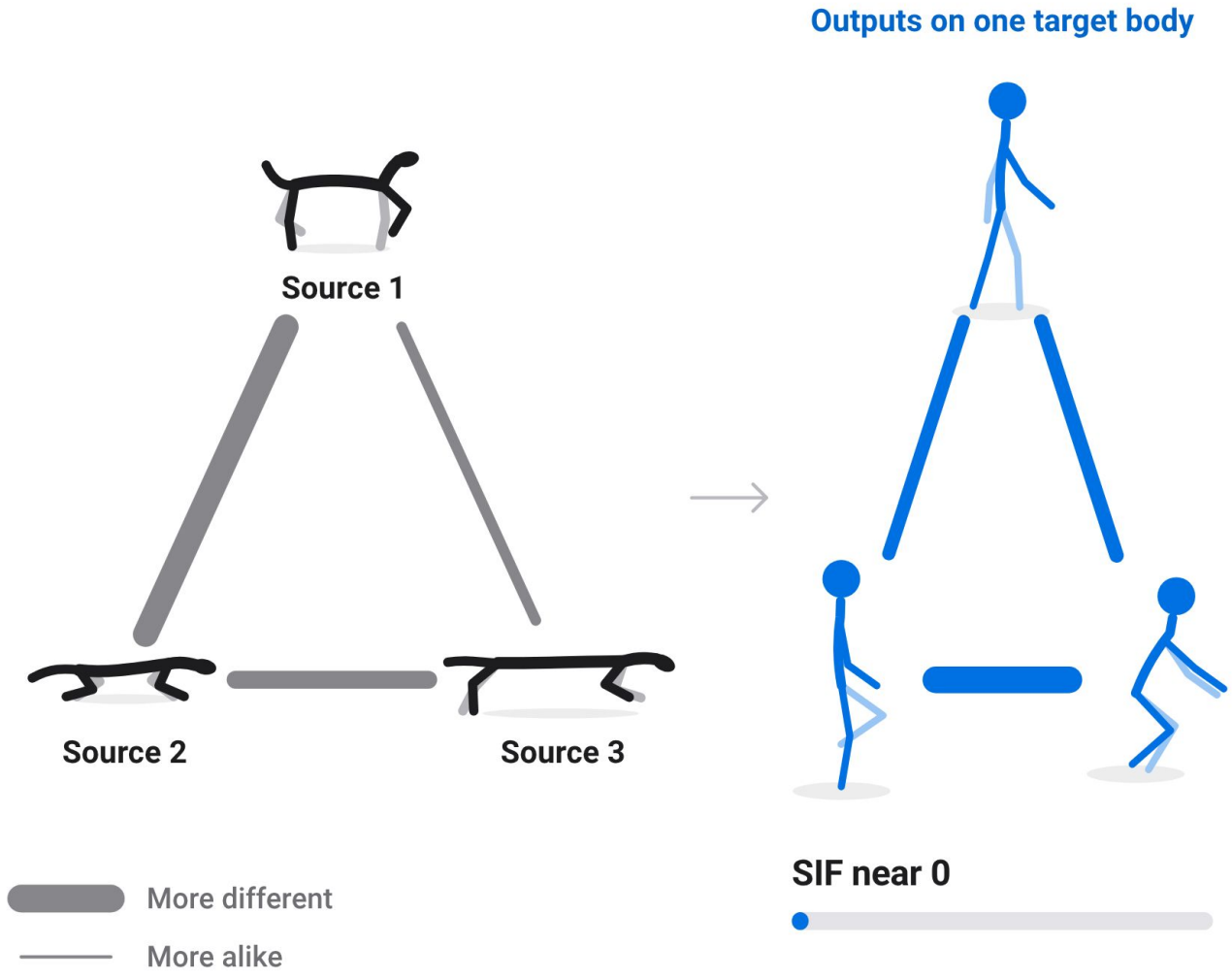
A line joins every pair. The thicker the line, the more the two clips differ.

A faithful model: the output lines match the source lines.

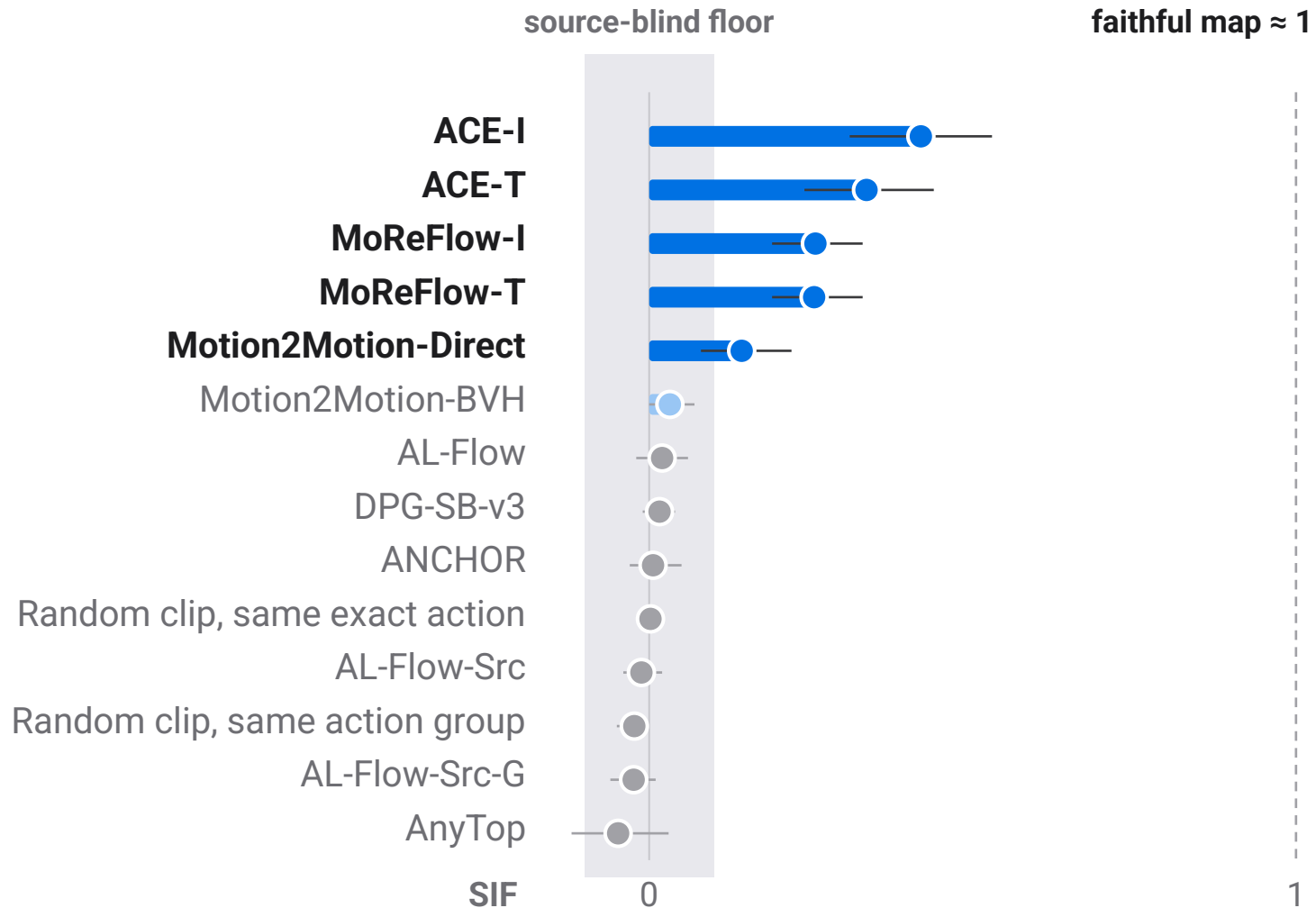
A model that ignores the source: its output lines no longer follow.

$$\rho_{\mathrm{SIF}} \;=\; \mathrm{corr}\Big(\{d_{\mathrm{src}}(i,j)\}_{i<j} \,,\;\; \{d_{\mathrm{out}}(i,j)\}_{i<j} \Big)$$

SIF correlates source and output distances; we report it with output variation, the ratio of their spreads.



Results on Truebones

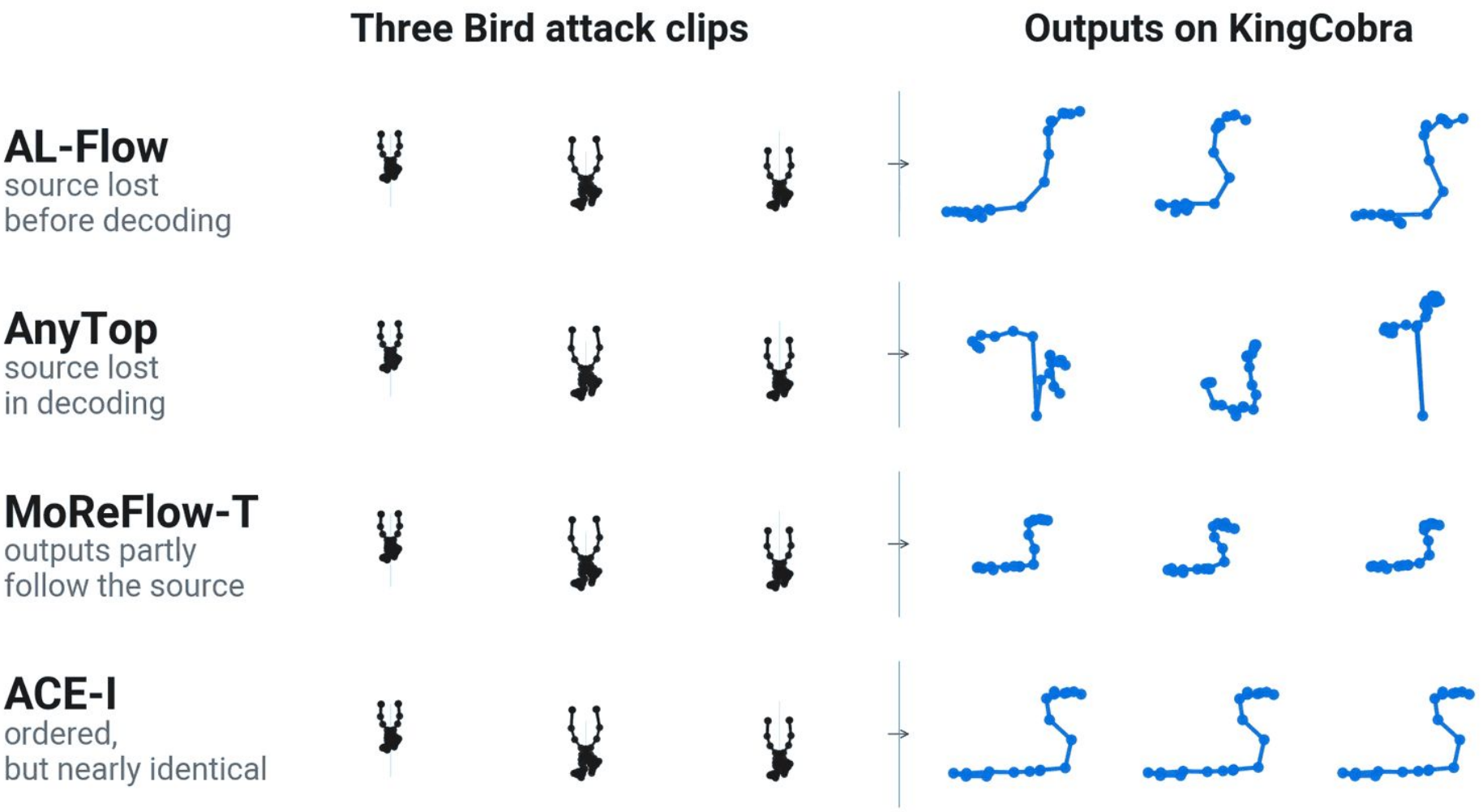


Nine of fourteen methods sit at or near the floor, including the label-only rules that matched the action-level score.

Above the floor, the signal is small: ACE's outputs vary only 2 to 3% as much as its sources.

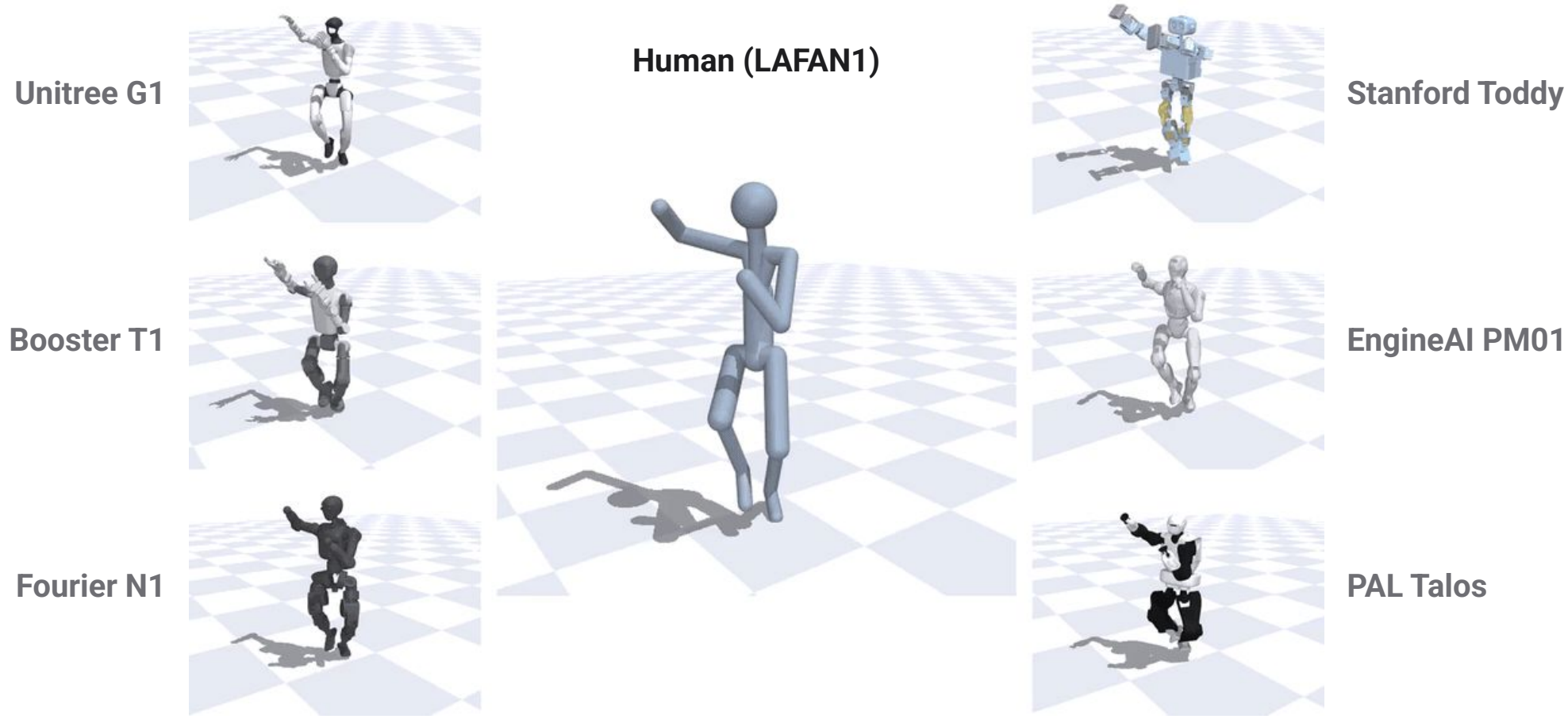
1,891 source–target–action triples; dots are raw SIF with 95% intervals that resample whole source skeletons. Paper Table 9.

Losing the Source



Bird → King cobra, attack. Three source clips; each row shows one method's three outputs.

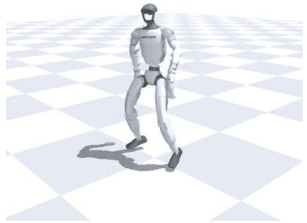
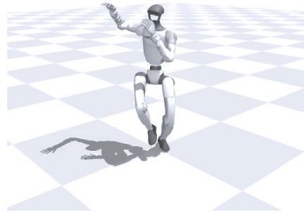
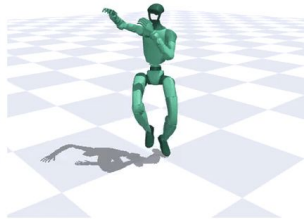
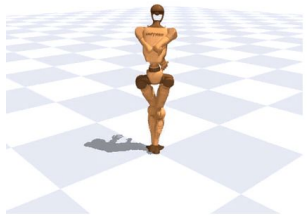
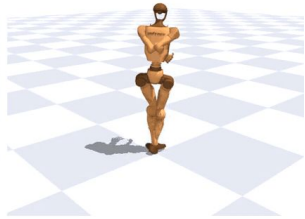
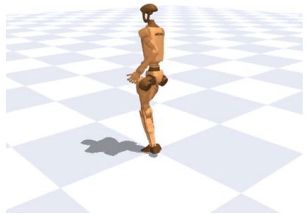
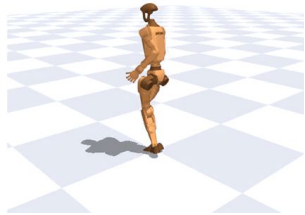
Human-to-Robot Study



A retargeting tool gives every human clip a robot version, which we treat as the true answer.

Human motion: LAFAN1 (Ubisoft La Forge, CC BY-NC-ND 4.0). Robot motion: GMR retargeting tool (MIT License).

Two Dancers on the Unitree G1

	Dancer 1	Dancer 2	difference between dancers kept
True retarget			reference
Trained on true pairs			99%
Trained on random pairs same routine, random partner			3.5%
Trained without pairs per-body latent matching			1.2%

26 LAFAN1 routines on the Unitree G1, relative to the true retarget (paper Table 29). The dancers are not in step, so part of what you see is timing.

Conclusion

1 **Standard objectives do not identify the source-conditioned map**

2 **Action-level scores cannot see this.** SIF, reported with output variation, can

3 **More action-labelled data does not recover the map; *correspondence* evidence does.** How little correspondence is enough is open

Project page

Paper

Code