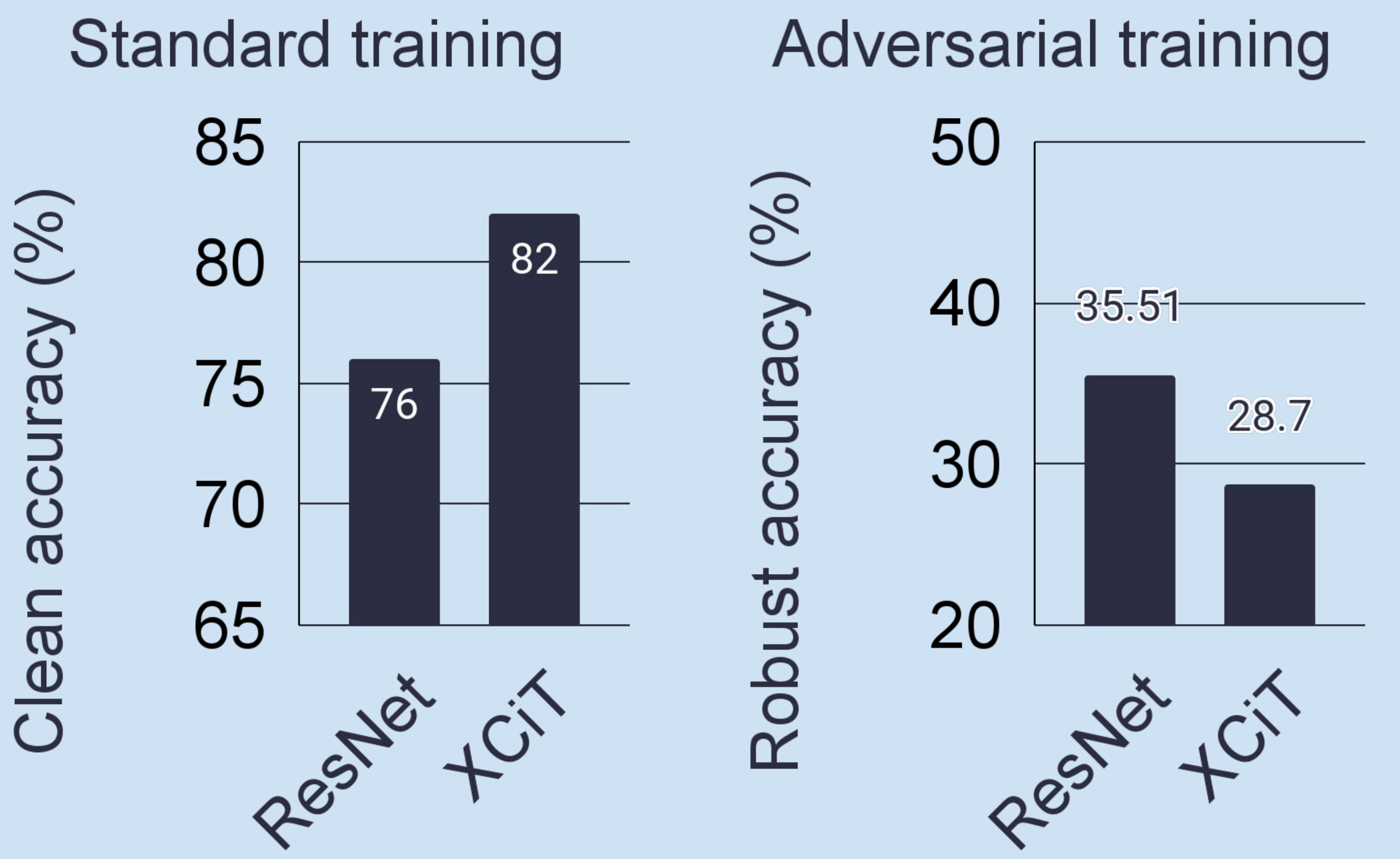
A *Light* Recipe to Train Robust Vision Transformers

Edoardo Debenedetti (ETH Zurich - EPFL), Vikash Sehwag (Princeton University), Prateek Mittal (Princeton University)

1/ The standard recipe for ViTs is suboptimal for adversarial training

Strong data augmentation and lower weight decay help in standard training, but hurt adversarial training

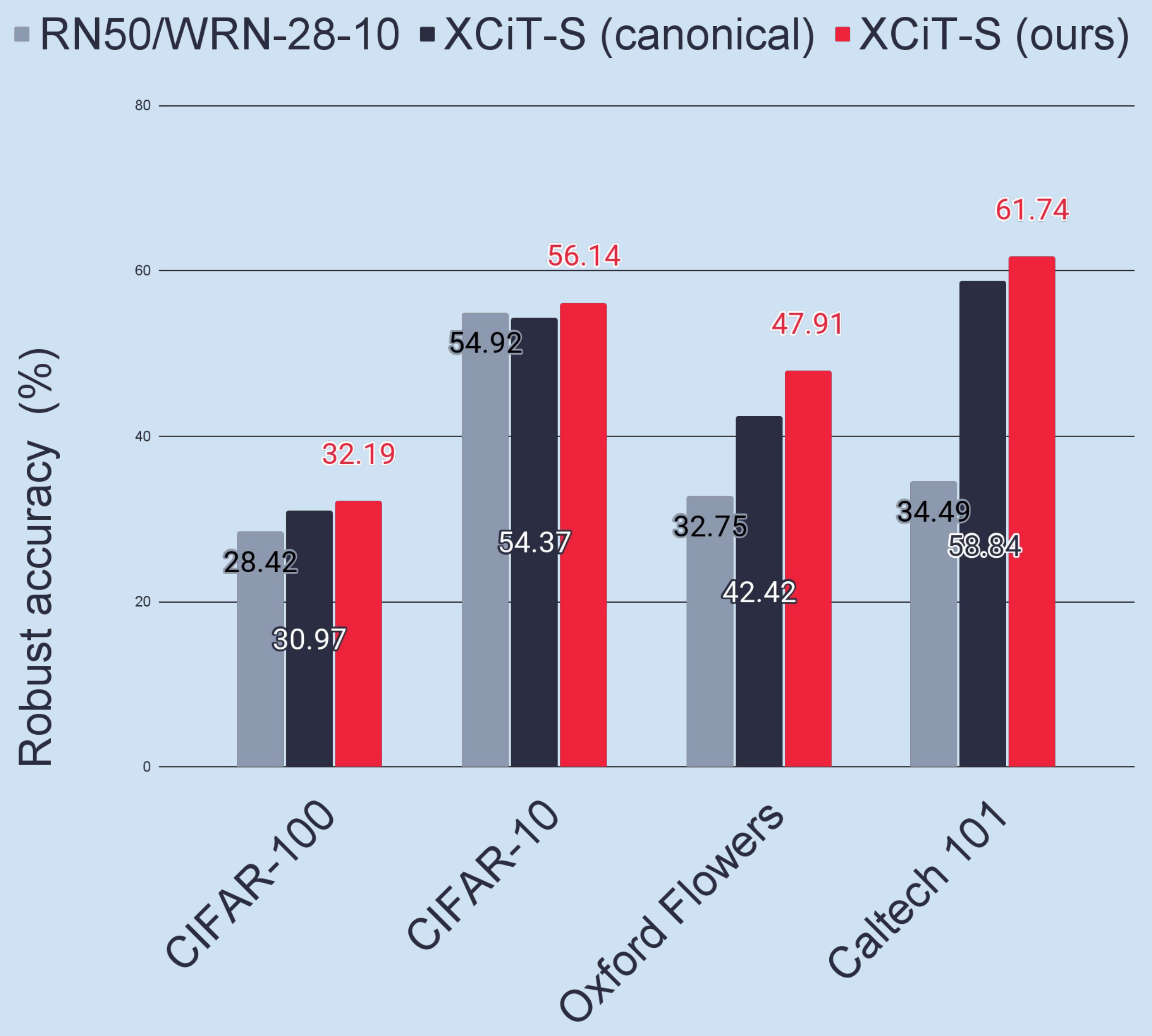


2/ Our "light" recipe

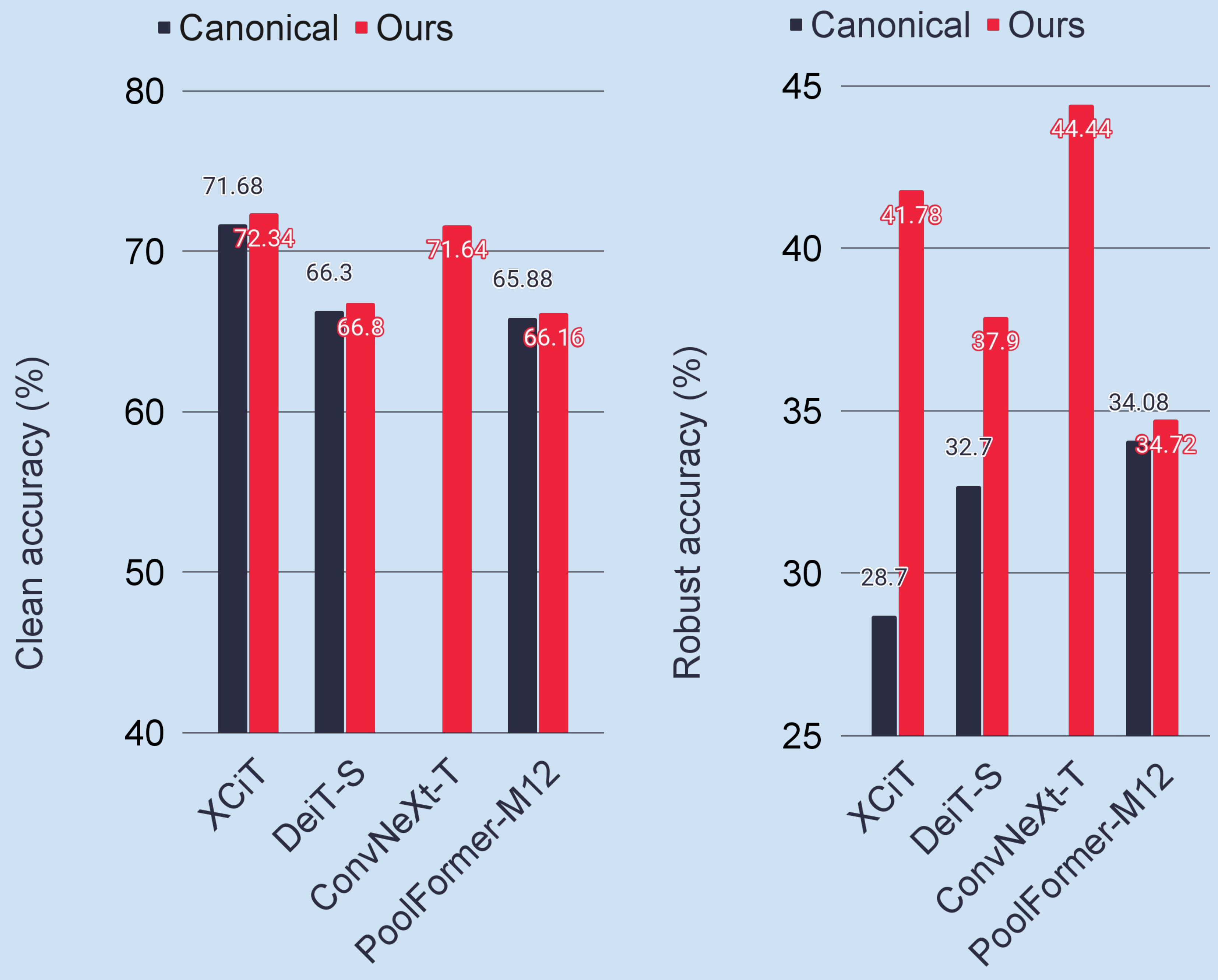
- 10 epochs linear ε-warmup
- basic data augmentation
- high weight decay

Feature	Accuracy	
	Clean	Robust
XCiT-S12	71.68	28.70
+ ε-warmup	71.98 (+0.30)	29.36 (+0.66)
+ Tuned data augmentation	71.70 (-0.28)	38.78 (+9.42)
+ Tuned weight decay	72.34 (+0.64)	41.78 (+3.00)

3/ The recipe generalizes to other datasets via fine-tuning …

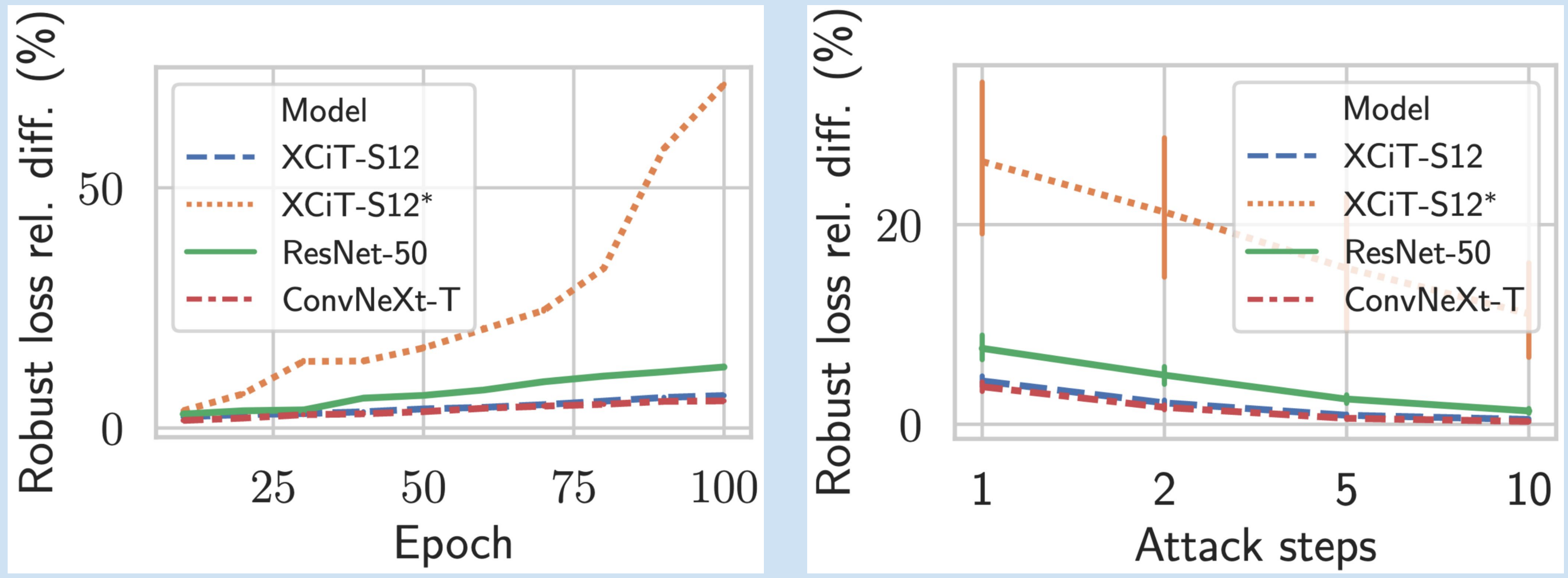


4/ … and to other architectures



5/ The recipe affects adversarial training's inner optimization

Attacking a model with few steps is easier for some architectures than for others, when trained with the right training recipe. This makes the resulting models more robust.



6/ Perceptual perturbations

We quantify that perturbations targeting more robust models are more aligned with perception