

Motivation: accuracy is not validity

Large sleep cohorts are now scored by models, not people.

- **Task.** Modality data $\rightarrow$ Wake / NREM / REM $\rightarrow$ sleep science.
- **Accuracy: solved.** Models match human scorers as often as scorers match each other, 80–85% (Danker-Hopfe et al., 2009).
- **Validity: not.** An accurate night can still be impossible: Wake $\rightarrow$ REM, one-epoch bouts, rapid alternation. Smoothing hides these and costs accuracy.
- **Problem.** Rule out the impossible, not the rare: rare transitions are diagnostic.

TVR and FI: two validity indicators

Read off predicted labels alone, no ground truth. $\mathcal{R}$ is the rare-transition set and

$$v_t = \mathbf{1}[(y_{t-1}, y_t) \in \mathcal{R}].$$

Modality	Rare Transitions	Rationale
EEG/EMG (3-state)	W $\rightarrow$ R, R $\rightarrow$ N	SOREMPs pathological
Cardiorespiratory	W $\rightarrow$ R, R $\rightarrow$ N	Same
Bioradar (3-state)	W $\rightarrow$ R, R $\rightarrow$ N	Same
Actigraphy (2-state)	—	None

Table 1. $\mathcal{R}$ by modality.

Experts score TVR = 2.1%; models 3.6–12.3%.

StageGuard: a plug-and-play constraint wrapper

A penalty during training, a constrained decoder at inference.

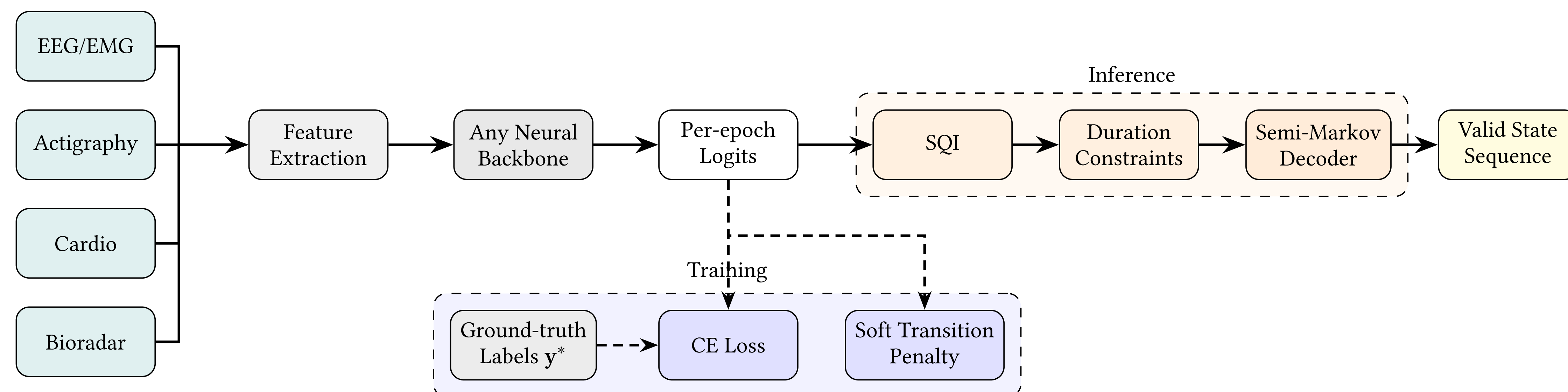


Figure 1. Wraps any backbone; nothing inside changes.

Training. Penalize probability mass on rare pairs:

$$\mathcal{L}_{\text{trans}} = \frac{1}{T-1} \sum_{t=2}^T \sum_{(s,s') \in \mathcal{R}} p_{\theta}(y_{t-1}=s \mid x_{t-1}) p_{\theta}(y_t=s' \mid x_t) \quad (3)$$

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{trans}}, \quad \lambda = 0.5 \quad (4)$$

Inference. Dwell counter, $D_{\text{max}}=10$: stages last a minimum time.

$$\tilde{A}_{(s,d),(s',d')} = \begin{cases} 0 & s'=s, d'=\min(d+1, D_{\text{max}}) \\ \log \epsilon_{ss'} & s' \neq s, d'=1, d \geq d_{\min}(s) \\ -\infty & \text{otherwise} \end{cases} \quad (5)$$

- **Discouraged, not forbidden.** Rare transitions cost $\log \epsilon_{ss'}$, $\epsilon=10^{-3}$, not $-\infty$: strong evidence recovers them ($\geq 91\%$ recall).
- **Labeled, not claimed.** Decoder and $d_{\min}$ exact; penalty and anti-alternation term not. +0.8 ms per recording (U-Sleep).

Results: six backbones $\times$ four modalities

Validity collapses on every pair; FI falls 53–64%, accuracy comes along.

Backbone	AccuSleep Mouse EEG/EMG			Sleep-Accel (Actigraphy)			SHHS			SLEEPBRL (Bioradar)		
	TVR↓	FI↓	Acc↑	TVR↓	FI↓	Acc↑	TVR↓	FI↓	Acc↑	TVR↓	FI↓	Acc↑
AccuSleep +Ours	8.7±0.7 0.6±0.1***	0.21±0.01 0.08±0.006***	89.3±0.7 91.0±0.6**	—	0.25±0.01 0.10±0.01***	84.1±0.7 87.4±0.6**	8.1±0.6 0.5±0.1***	0.22±0.01 0.09±0.007***	76.8±0.8 80.5±0.7**	12.3±0.9 0.8±0.2***	0.28±0.02 0.12±0.01***	68.4±1.1 73.2±1.0**
DeepSleepNet +Ours	6.4±0.5 0.4±0.1***	0.17±0.01 0.07±0.005***	90.4±0.6 91.5±0.6**	—	0.21±0.01 0.09±0.006***	85.7±0.6 85.6±0.6**	6.0±0.5 0.3±0.06***	0.18±0.01 0.08±0.005***	78.9±0.7 81.3±0.6**	10.1±0.8 0.5±0.1***	0.24±0.02 0.10±0.01***	70.6±1.0 74.5±0.9**
SeqSleepNet +Ours	5.6±0.5 0.3±0.07***	0.16±0.01 0.07±0.004***	90.8±0.6 91.7±0.5**	—	0.19±0.01 0.08±0.005***	86.2±0.6 88.6±0.5**	5.3±0.4 0.3±0.06***	0.17±0.01 0.07±0.004***	79.4±0.7 81.7±0.6**	9.2±0.7 0.5±0.1***	0.23±0.01 0.10±0.01***	71.4±0.9 75.2±0.8**
AttnSleep +Ours	4.5±0.4 0.3±0.06***	0.14±0.01 0.06±0.004***	91.1±0.5 91.9±0.5**	—	0.18±0.01 0.08±0.005***	86.5±0.5 88.8±0.5**	4.2±0.3 0.2±0.05***	0.15±0.01 0.07±0.004***	79.9±0.6 82.0±0.5**	7.8±0.6 0.4±0.1***	0.21±0.01 0.09±0.01***	72.0±0.8 75.8±0.8**
SleepTransformer +Ours	4.9±0.4 0.3±0.1***	0.15±0.01 0.06±0.003***	91.3±0.5 91.2±0.5	—	0.17±0.01 0.07±0.004***	86.9±0.5 89.0±0.4**	4.6±0.4 0.2±0.05***	0.16±0.01 0.06±0.003***	80.3±0.6 82.3±0.5**	8.5±0.6 0.3±0.1***	0.22±0.01 0.08±0.004***	72.7±0.8 76.3±0.7**
U-Sleep +Ours	3.8±0.3 0.2±0.05***	0.13±0.01 0.05±0.003***	91.6±0.4 92.2±0.4**	—	0.16±0.01 0.07±0.004***	87.2±0.5 89.3±0.4**	3.6±0.3 0.1±0.04***	0.14±0.01 0.06±0.003***	80.8±0.5 82.6±0.5**	6.7±0.5 0.3±0.05***	0.19±0.01 0.08±0.003***	73.4±0.7 77.0±0.7**

Table 2. Leave-one-subject-out means $\pm$ std; **bold** best per column, “—” = TVR undefined for 2-state actigraphy. Every TVR and FI difference significant (Wilcoxon, Bonferroni; ** $p<0.01$, *** $p<0.001$).

Downstream endpoints and ablations

We optimize none of these endpoints; they improve anyway.

Statistic	Baseline	+Ours	Improv.
Total Sleep Time (min)	18.4±2.1	7.2±1.0***	61%
Sleep Efficiency (%)	3.8±0.5	1.4±0.2***	63%
REM Latency (min)	21.3±3.3	8.7±1.6***	59%
WASO (min)	16.2±2.5	5.8±1.2***	64%
Mean NREM Bout (epochs)	4.2±0.6	1.1±0.2***	74%
Mean REM Bout (epochs)	2.4±0.4	0.6±0.1***	75%
Awakenings (count)	15.1±2.3	3.2±0.6***	79%

U-Sleep, mouse EEG/EMG. Table 3: endpoint error. Table 4: one component per row.

Constraint-effect analysis

Three views of the same effect, on AccuSleep mouse EEG/EMG.

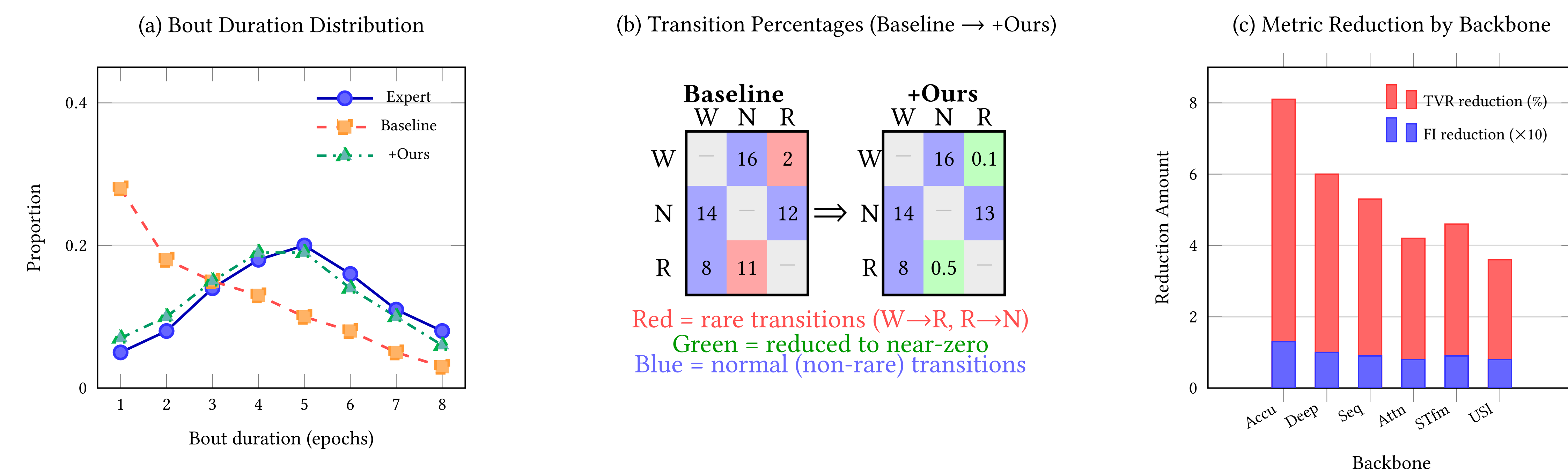
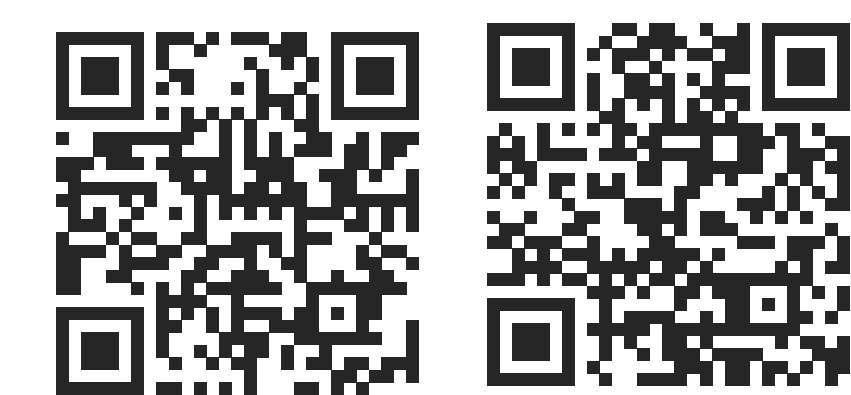


Figure 2. (a) Bouts match the expert. (b) Rare transitions collapse. (c) Both metrics fall everywhere.

Takeaways & links

- **Discourage, don’t forbid.** A ban destroys the rare events that matter.
- **Validity separates from accuracy.** TVR 3.6–12.3% $\rightarrow$ 0.1–0.8%, accuracy within 0.1 pt.
- **Scope.** Healthy adults and mice; not narcolepsy, RBD or neonatal sleep.



Code

Paper