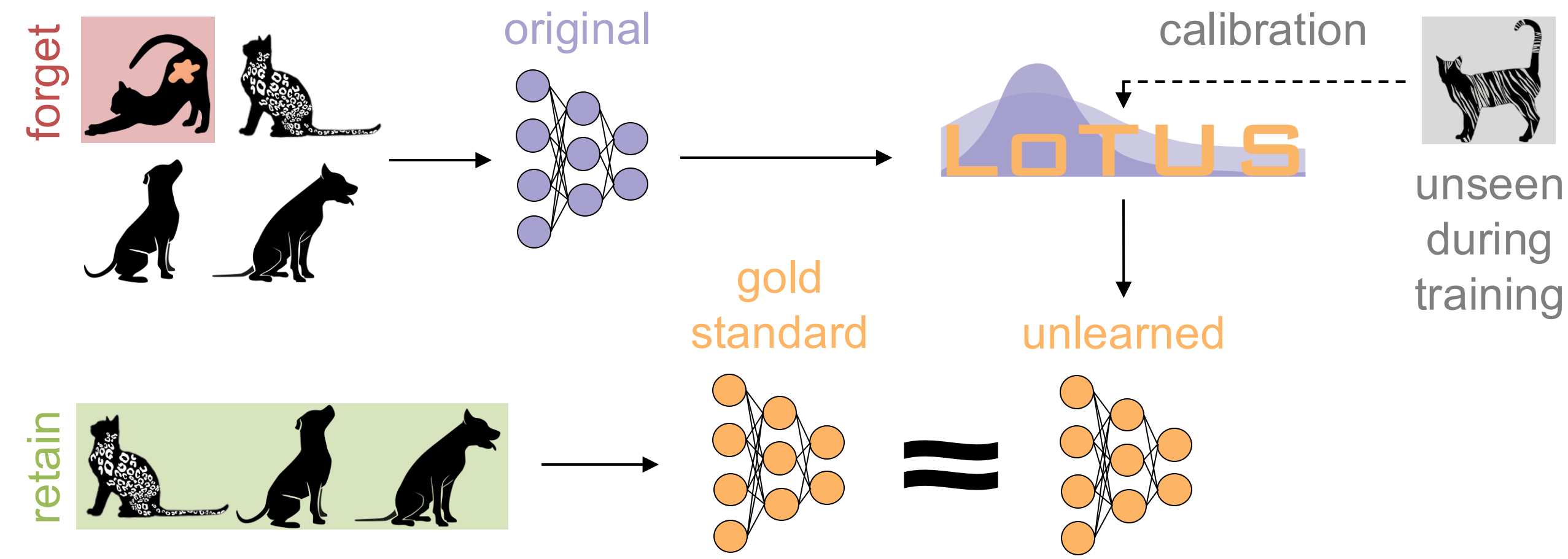
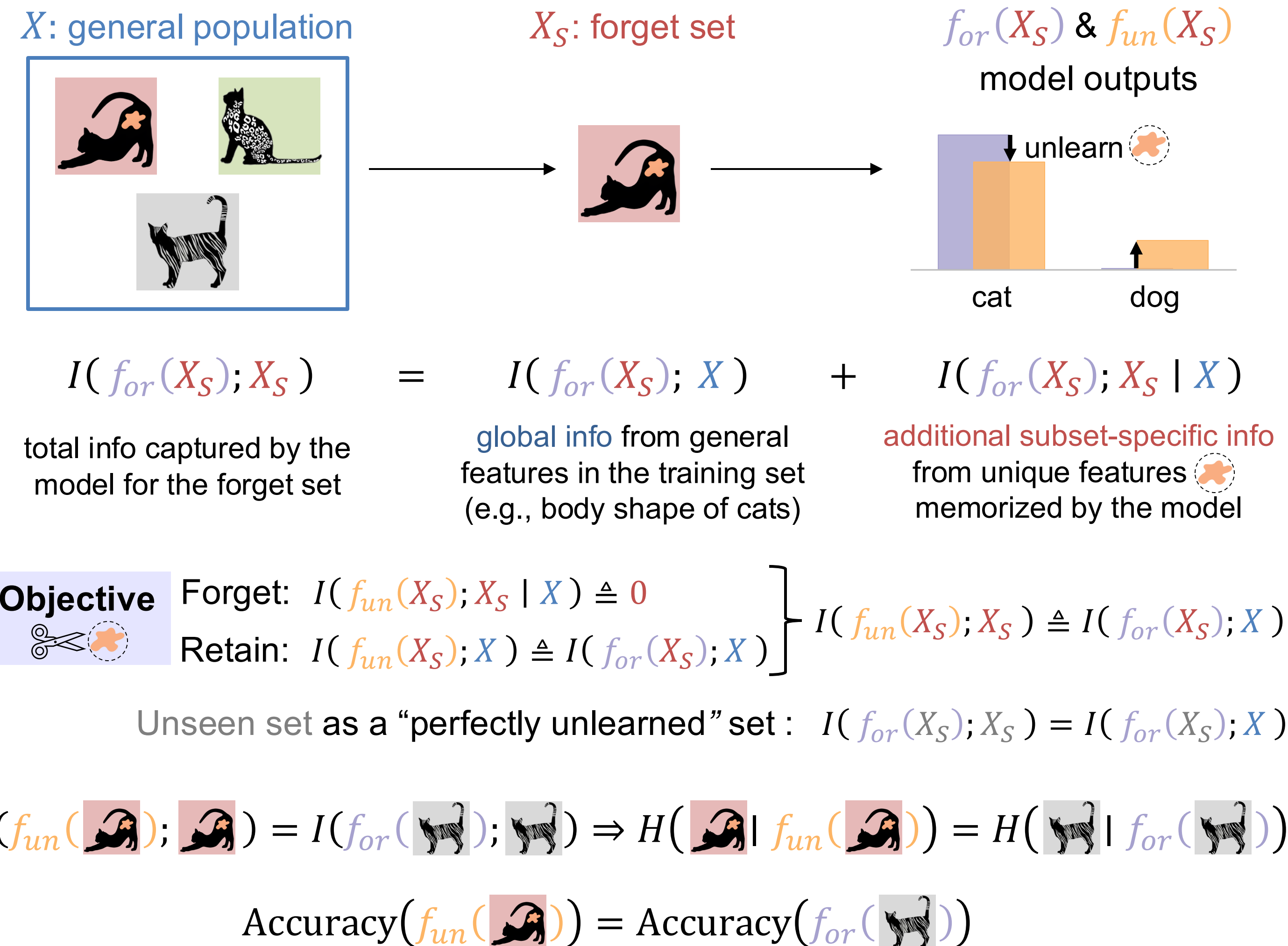


Machine Unlearning



Information-Theoretic Framework

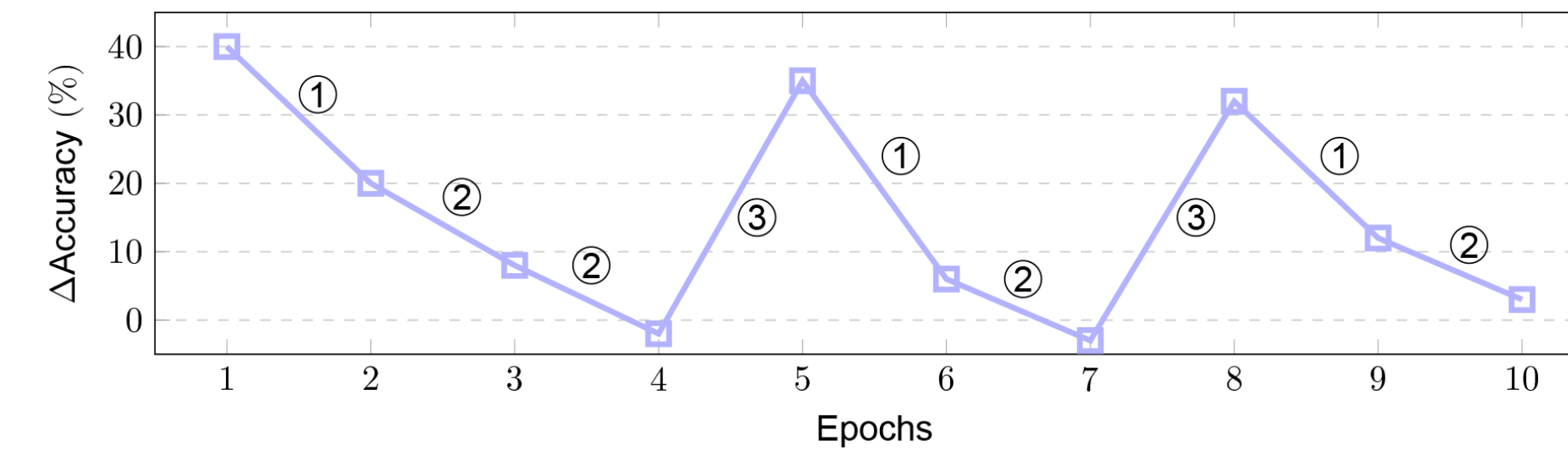


Logits Tempering Unlearning Strategy

$$\tau_d = \exp \left(\alpha \left(\text{Accuracy}(f_{un}(\text{cat})) - \text{Accuracy}(f_{or}(\text{dog})) \right) \right)$$

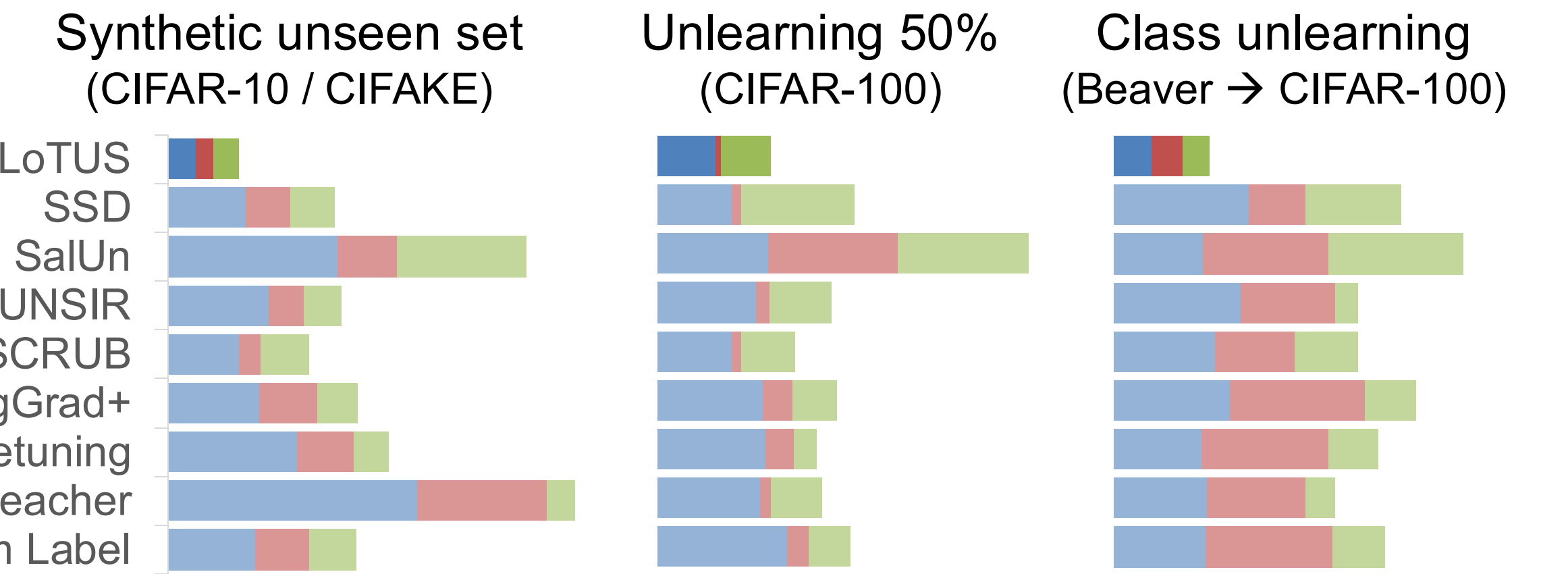
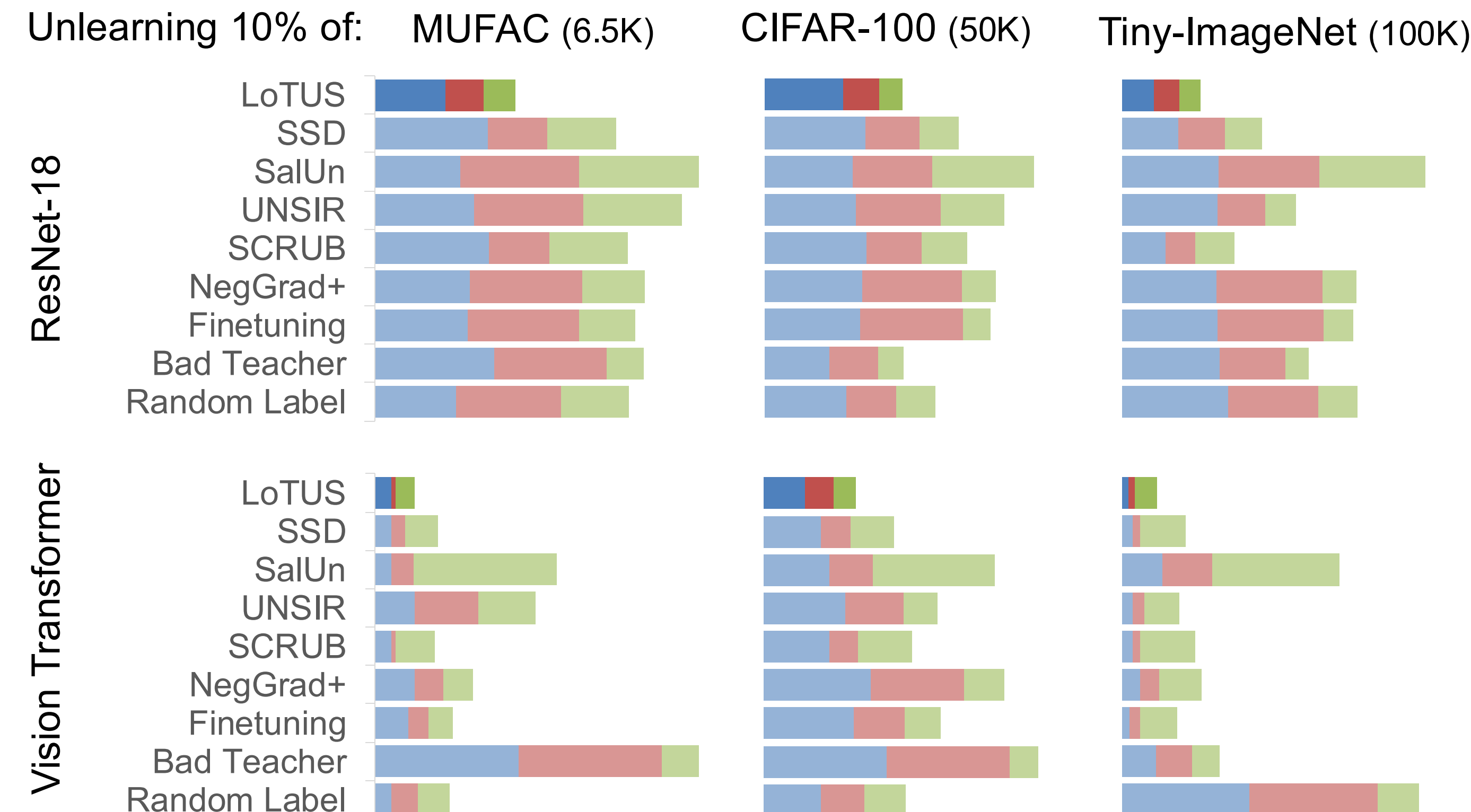
$$\text{Forget loss: } gs(f_{or}(\text{cat}), \tau_d) \odot \log s(f_{un}(\text{cat}))$$

$$\text{Retain loss: } gs(f_{or}(\text{dog}), \tau \rightarrow 0^+) \odot \log s(f_{un}(\text{dog}))$$



Results

- Average Gap** (↓): Balance between forgetting and retention
- JSD** (↓): Unlearning efficacy & Resilience to the Streisand Effect
- Runtime Estimation** (↓): Unlearning efficiency



Large-Scale Real-World Benchmarking

The **gold standard** model is **not available**

- Pre-trained ViT
- ImageNet-1K (1.2M)
- Limited data access

Method	RF-JSD $\times 1e4$ (↓)	Time (↓)	Retain Acc.	MIA Acc.
Original	1.22 $\pm$ 0.01	(pre-trained)	0.94 $\pm$ 0.00	0.71 $\pm$ 0.00
Finetuning	2.22 $\pm$ 0.02	16.24 $\pm$ 0.03	0.97 $\pm$ 0.00	0.78 $\pm$ 0.00
NegGrad+	2.17 $\pm$ 0.02	18.10 $\pm$ 0.03	0.97 $\pm$ 0.00	0.80 $\pm$ 0.00
Rnd Labeling	1.80 $\pm$ 0.09	19.37 $\pm$ 0.03	0.95 $\pm$ 0.01	0.74 $\pm$ 0.01
Bad Teacher	3.16 $\pm$ 3.25	11.66 $\pm$ 0.03	0.77 $\pm$ 0.21	0.52 $\pm$ 0.18
SCRUB	1.24 $\pm$ 0.01	24.49 $\pm$ 0.03	0.94 $\pm$ 0.00	0.71 $\pm$ 0.00
SSD	1.23 $\pm$ 0.01	22.61 $\pm$ 0.10	0.94 $\pm$ 0.00	0.71 $\pm$ 0.00
UNSIR	2.54 $\pm$ 0.03	33.12 $\pm$ 0.03	0.99 $\pm$ 0.00	0.77 $\pm$ 0.01
SalUn	1.83 $\pm$ 0.03	59.27 $\pm$ 0.37	0.95 $\pm$ 0.00	0.74 $\pm$ 0.01
LoTUS	1.11$\pm$0.01	10.72$\pm$0.01	0.94$\pm$0.00	0.61$\pm$0.01

Novel Metric: Retrain-Free Jensen-Shannon Divergence

$$\left. \begin{aligned} JSD(f_{un}(\text{cat}) || f_{gold}(\text{cat})) \\ RF-JSD(f_{un}(\text{cat}) || f_{or}(\text{dog})) \end{aligned} \right\} \text{Highly correlated}$$

Contributions

- ✓ **Information-Theoretic Framework** for formalizing Machine Unlearning
- ✓ **LoTUS**: Scalable and effective entropy-based unlearning strategy
- ✓ **RF-JSD**: Evaluation metric for large-scale and real-world benchmarking



Blog
Code
Paper
Video