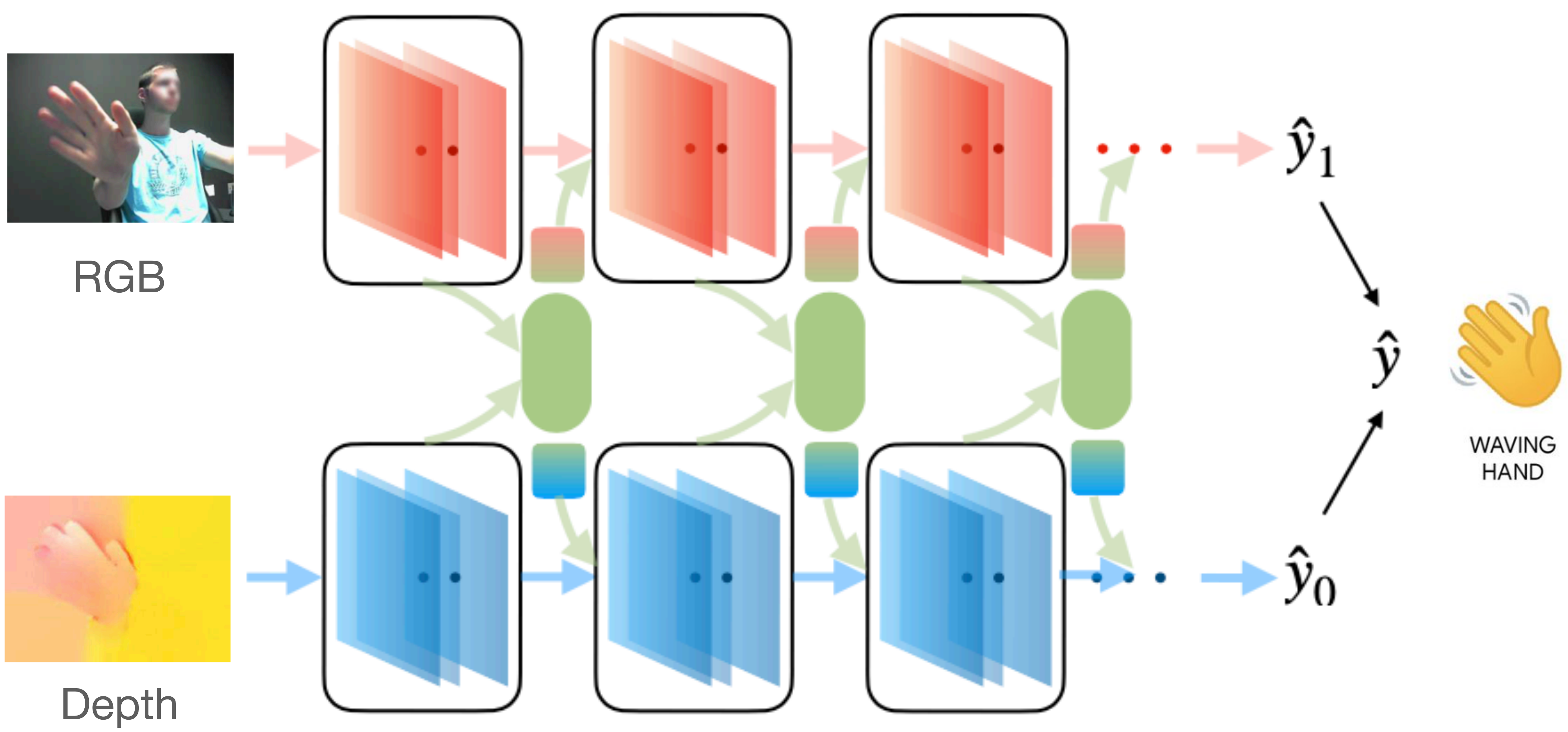
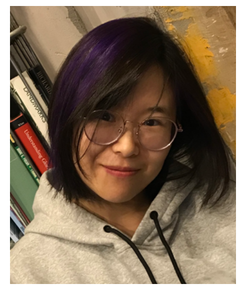
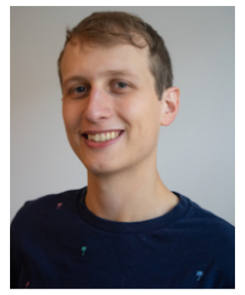
# Characterizing and Overcoming the Greedy Nature of Learning in Multi-Modal Deep Neural Networks

*Dynamic Hand Gestures Recognition*




**Nan Wu**, Stanisław Jastrzębski, Kyunghyun Cho and Krzysztof J. Geras.

NYU DATA SCIENCE

CAI²R

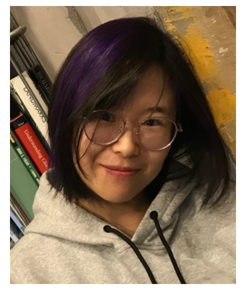
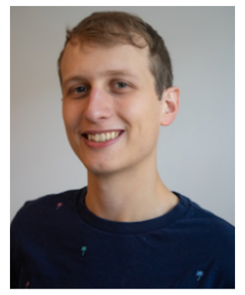
# Characterizing and Overcoming the Greedy Nature of Learning in Multi-Modal Deep Neural Networks

**Do multi-modal DNNs attend to all modalities?**

**Can we make multi-modal DNNs utilize all modalities?**

**Does better utilization of all modalities imply better generalization?**

**Nan Wu**, Stanisław Jastrzębski, Kyunghyun Cho and Krzysztof J. Geras.

NYU DATA SCIENCE

CAI²R

# Metric 1: Conditional utilization rate

$$\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1})$$

$$\hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1})$$

# Metric 1: Conditional utilization rate

$$\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}'_0 = f'_0(\boldsymbol{x}_{m_0})$$

$$\hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}'_1 = f'_1(\boldsymbol{x}_{m_1}).$$

# Metric 1: Conditional utilization rate

$$\boxed{\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}'_0 = f'_0(\boldsymbol{x}_{m_0})}$$

$$\hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}'_1 = f'_1(\boldsymbol{x}_{m_1}).$$

$$\boxed{\boldsymbol{u}(m_1|m_0) = \frac{A(f_0) - A(f'_0)}{A(f_0)},}$$

$$\boldsymbol{u}(m_0|m_1) = \frac{A(f_1) - A(f'_1)}{A(f_1)},$$

**The relative change in accuracy between the two models:**

**- one using all modalities,**

**- the other using only one.**

# Metric 1: Conditional utilization rate

$$\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}_0' = f_0'(\boldsymbol{x}_{m_0})$$

$$\hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}_1' = f_1'(\boldsymbol{x}_{m_1}).$$

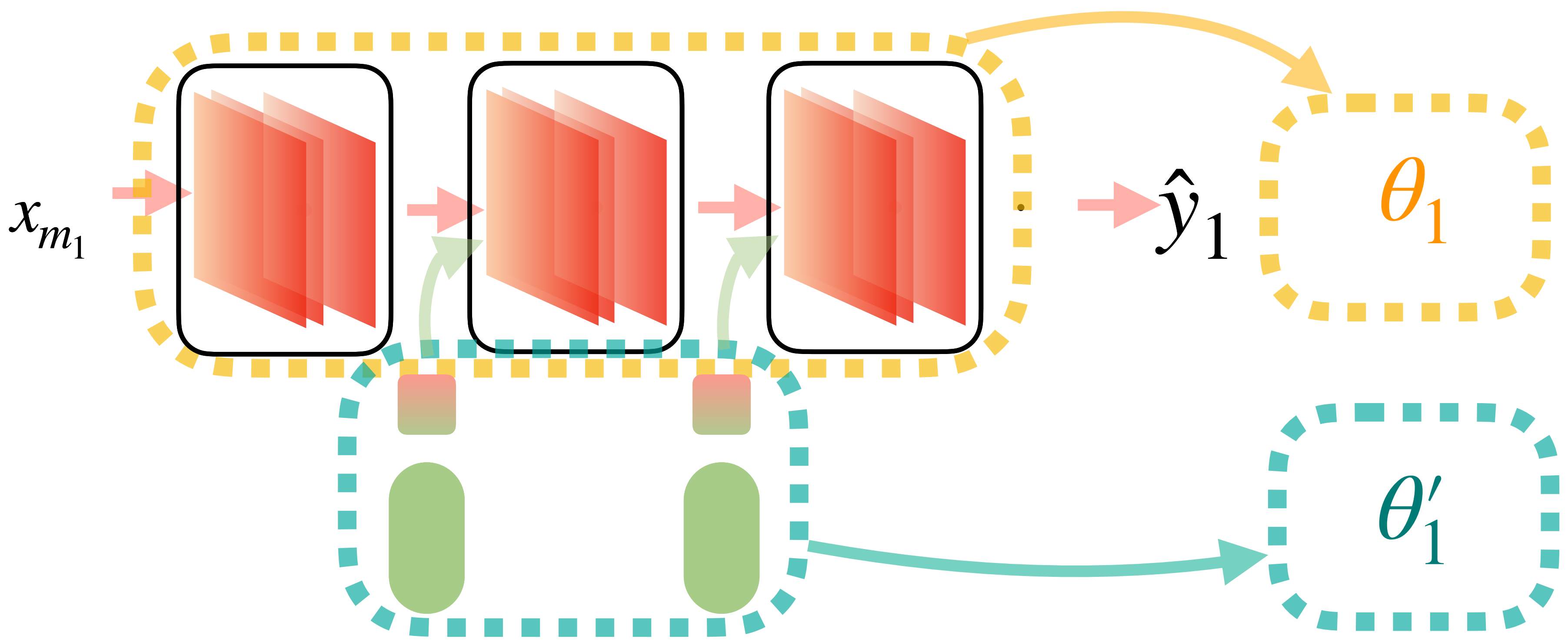
$$\boldsymbol{u}(m_1|m_0) = \frac{A(f_0) - A(f_0')}{A(f_0)},$$

$$\boldsymbol{u}(m_0|m_1) = \frac{A(f_1) - A(f_1')}{A(f_1)},$$

**The relative change in accuracy between the two models:**

- **one using all modalities,**
- **the other using only one.**

# Metric 2: Conditional learning speed



# Metric 1: Conditional utilization rate

$$\boxed{\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}'_0 = f'_0(\boldsymbol{x}_{m_0})}$$

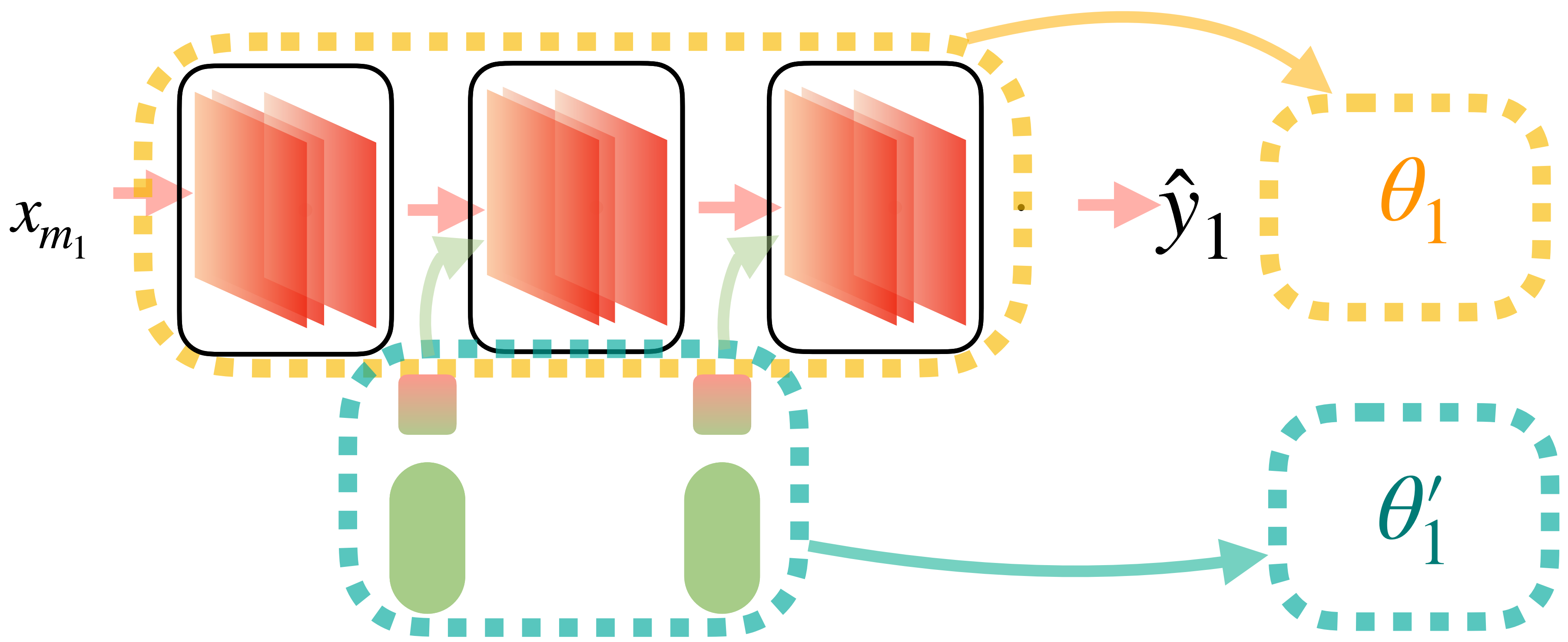
$$\hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}) \qquad \hat{y}'_1 = f'_1(\boldsymbol{x}_{m_1}).$$

$$\boxed{\boldsymbol{u}(m_1|m_0) = \frac{A(f_0) - A(f'_0)}{A(f_0)},}$$

$$\boldsymbol{u}(m_0|m_1) = \frac{A(f_1) - A(f'_1)}{A(f_1)},$$

**The relative change in accuracy between the two models:**
- **one using all modalities,**
- **the other using only one.**

# Metric 2: Conditional learning speed

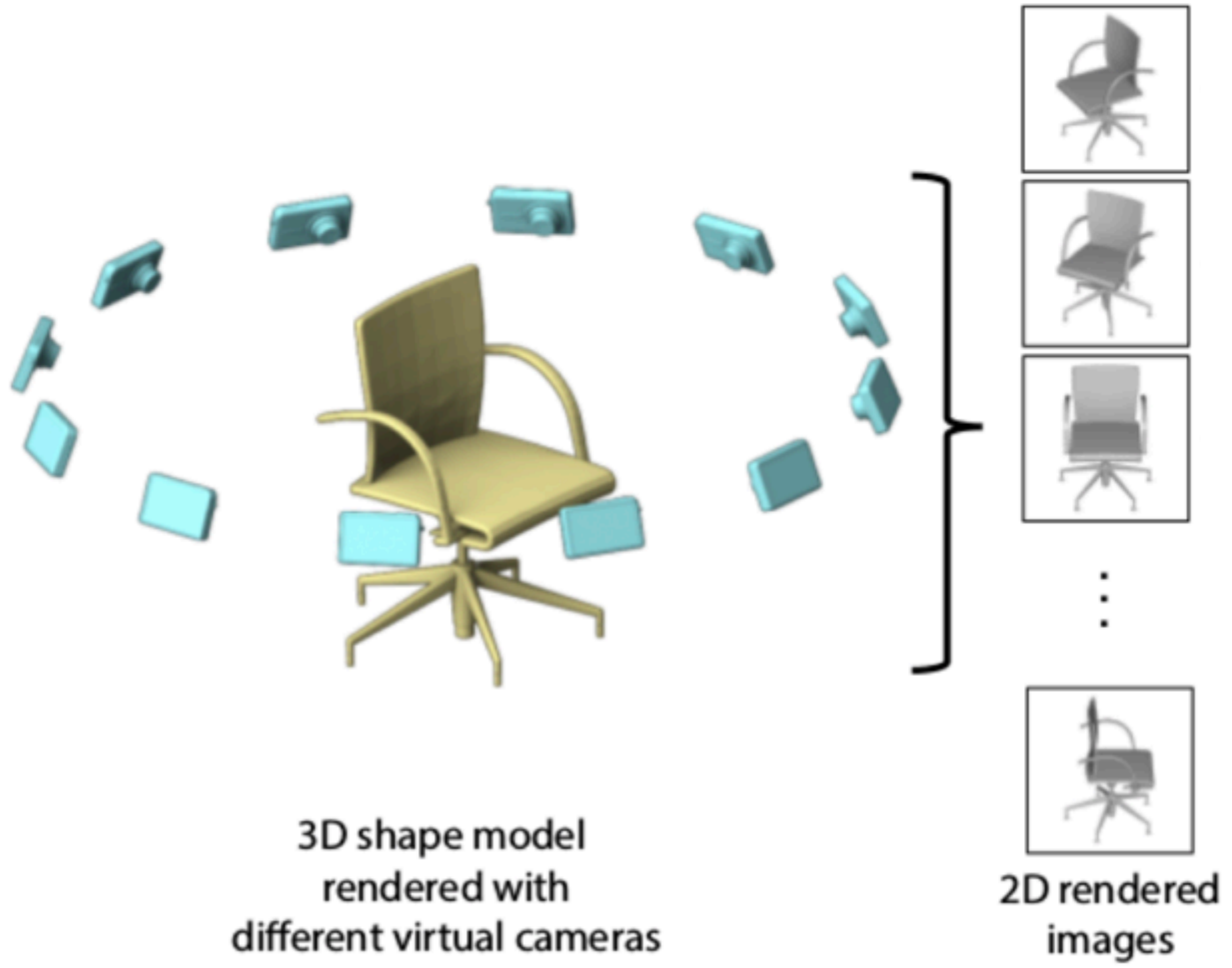



$$\boldsymbol{s}(m_0|m_1;t) = \log \frac{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}'_1;i)}{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}_1;i)},$$

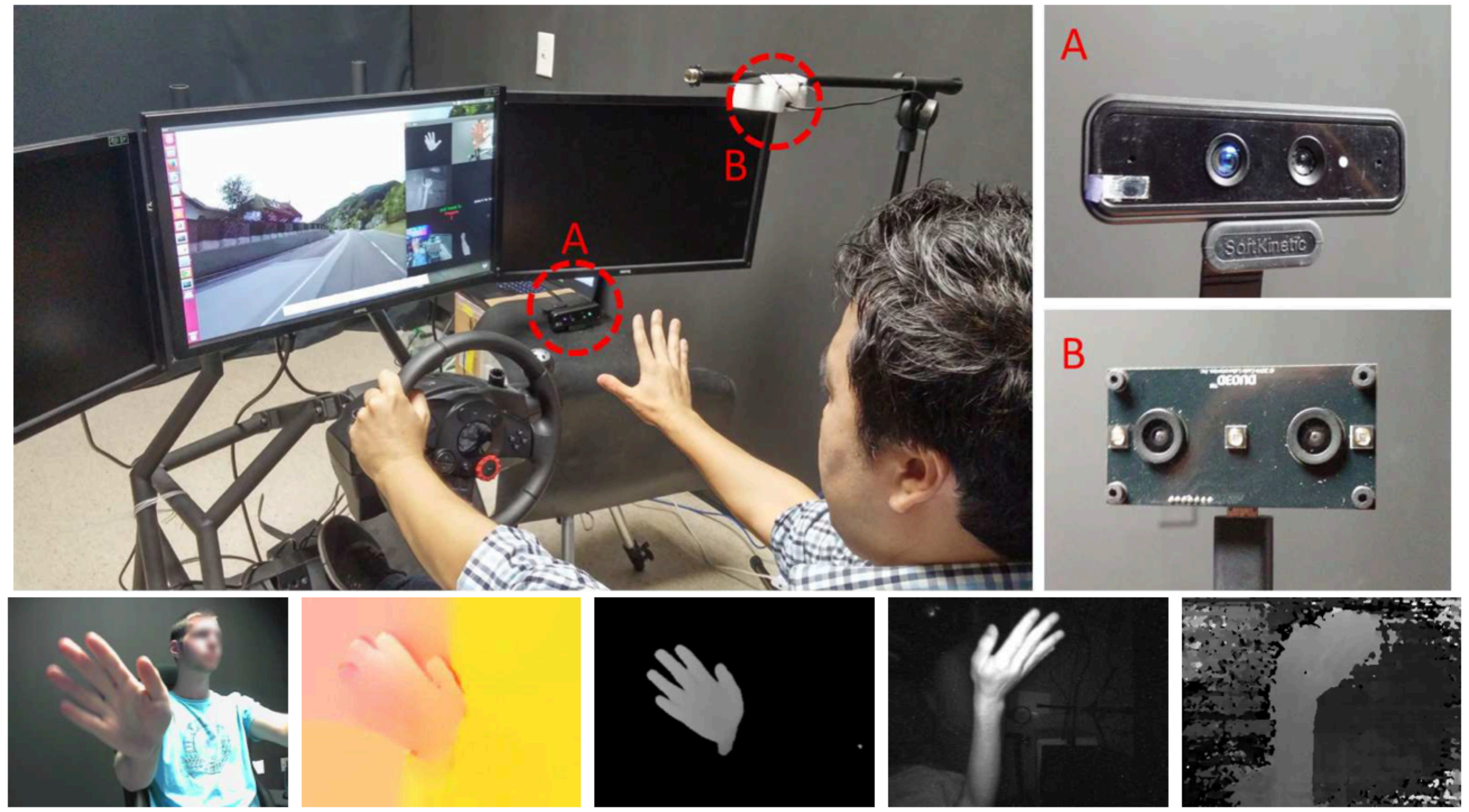
$$\mu(\boldsymbol{\theta};i) = ||\boldsymbol{G}||_2^2/||\boldsymbol{\theta}_{(i)}||_2^2$$

**the log-ratio between the learning speed of**
- **the uni-modal branch and**
- **the corresponding fusion components.**

# Observations: Imbalanced learning between modalities




(a) ModelNet40

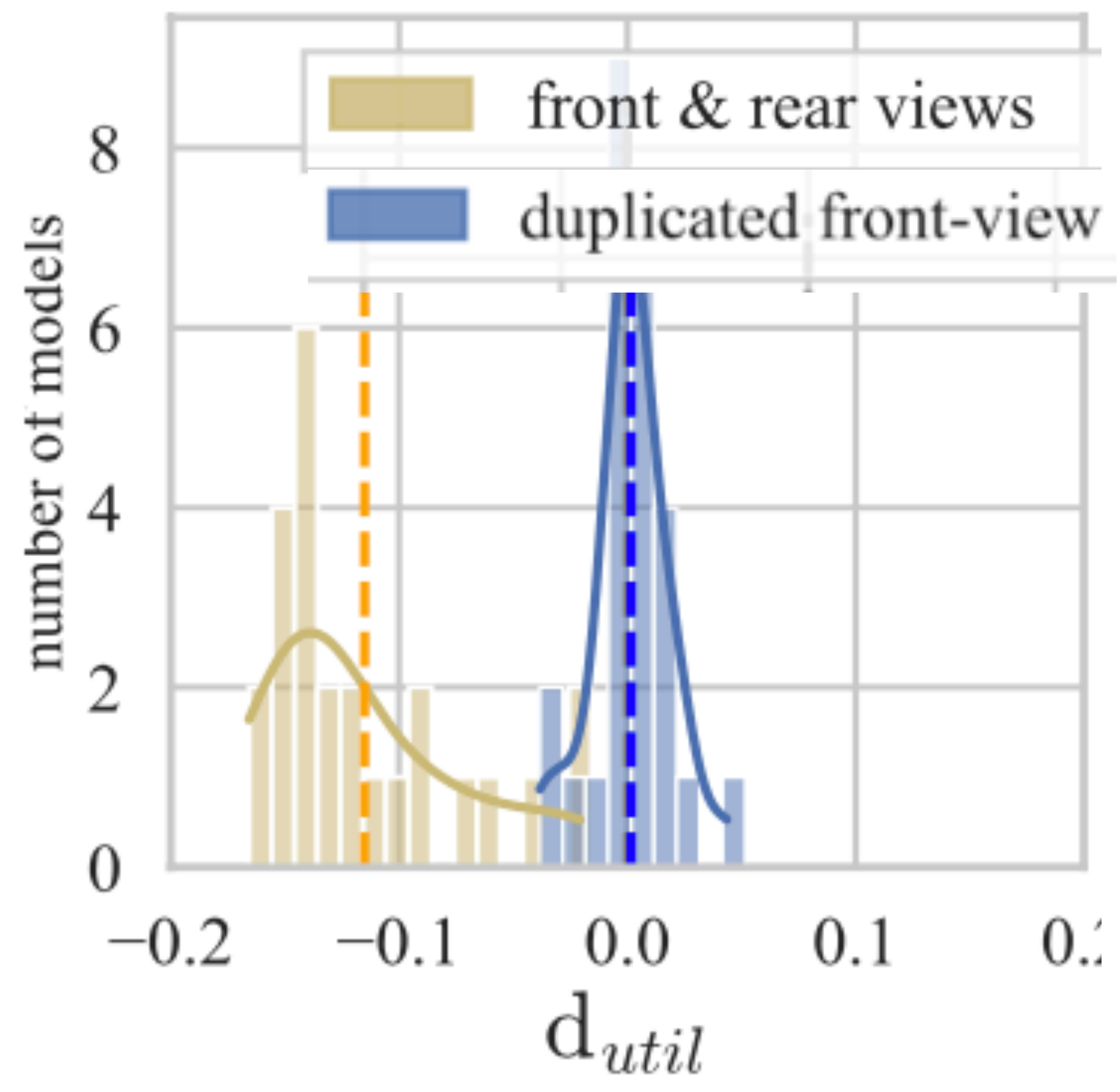



(b) NVGesture

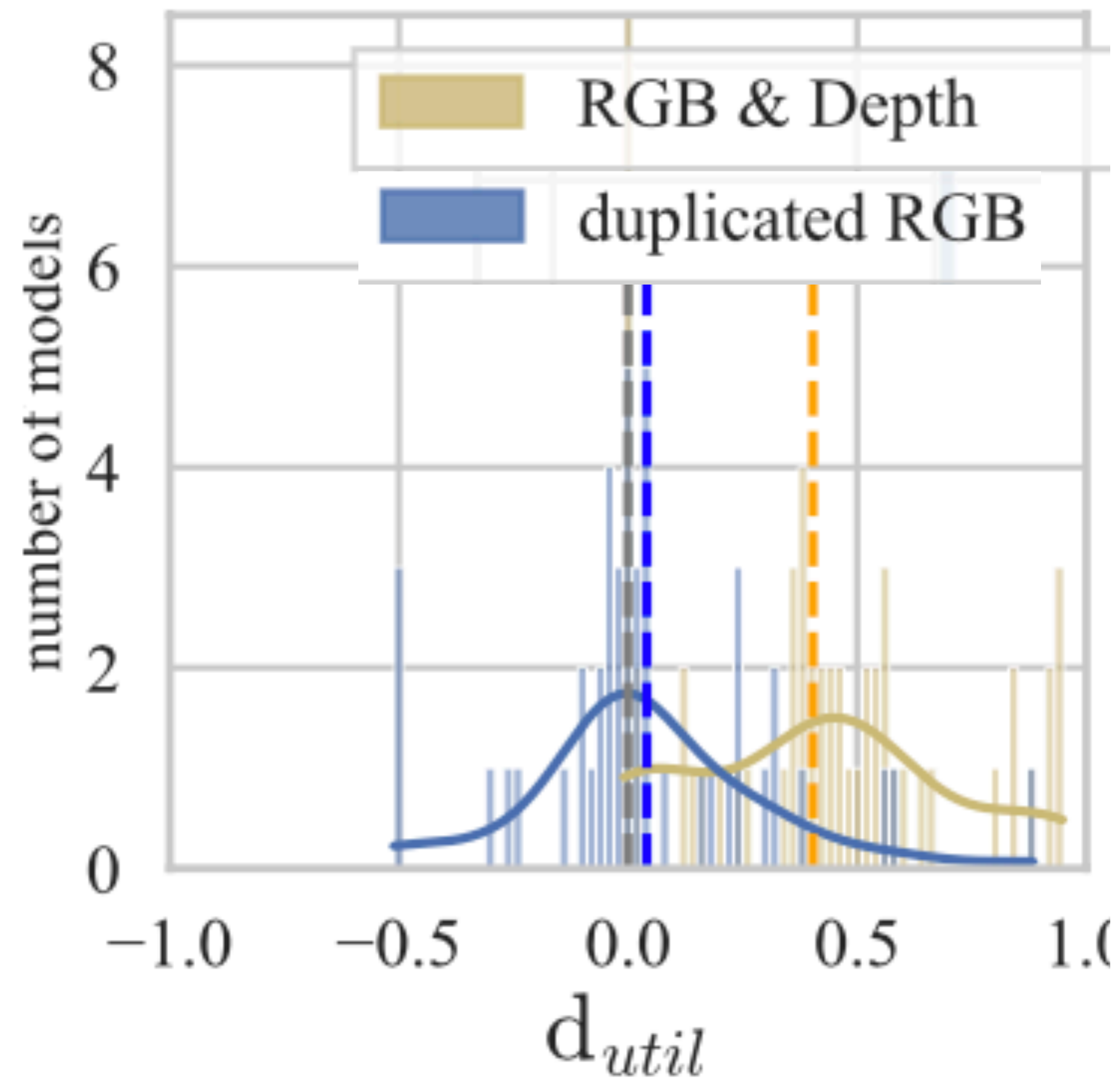
$$\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1)$$

$$\mathrm{d}_{\mathrm{speed}}(f;t) = \boldsymbol{s}(m_1|m_0;t) - \boldsymbol{s}(m_0|m_1;t)$$

# Observations: Imbalanced learning between modalities




(a) ModelNet40

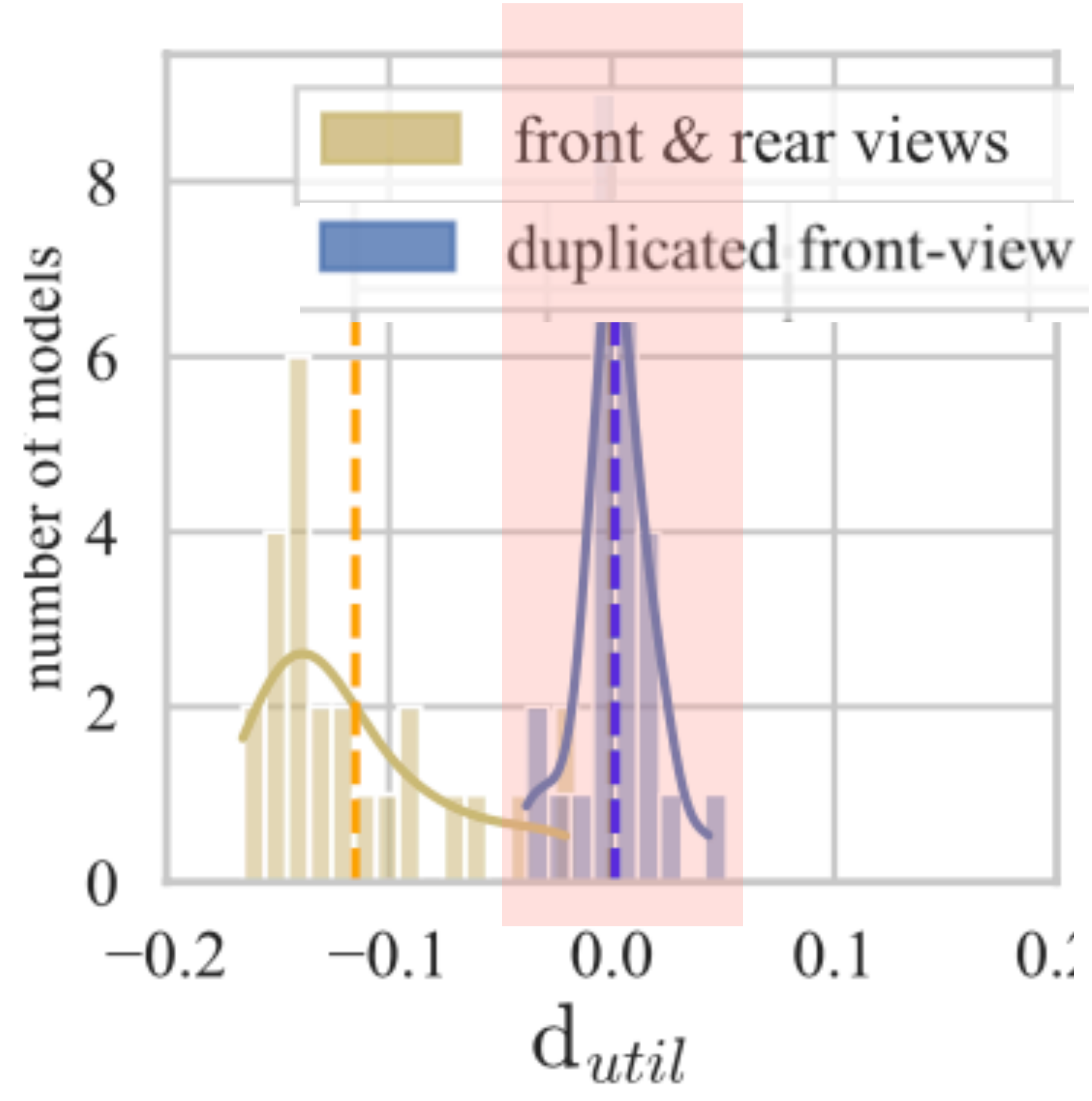



(b) NVGesture

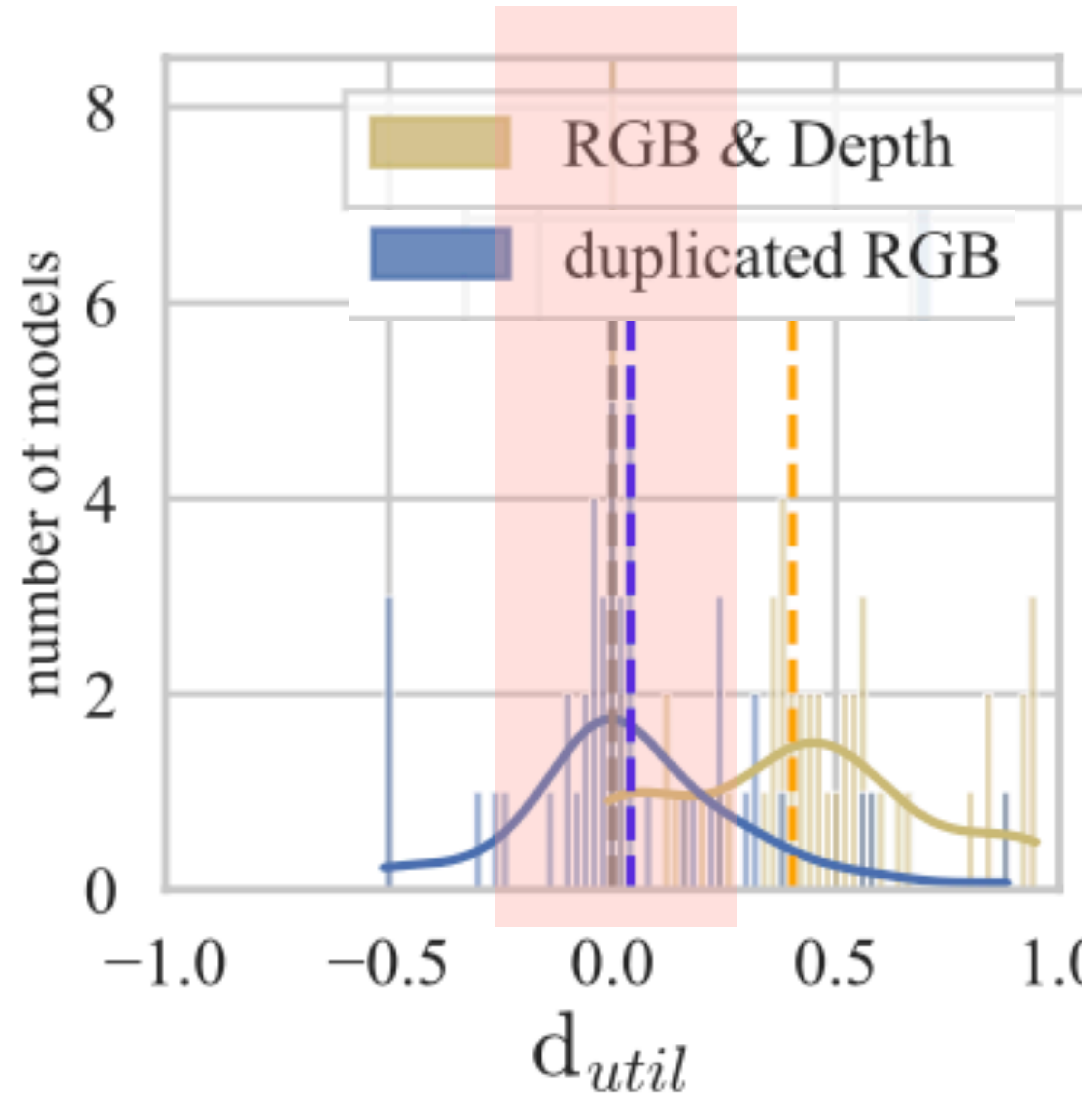
$$\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1)$$

$$\mathrm{d}_{\mathrm{speed}}(f;t) = \boldsymbol{s}(m_1|m_0;t) - \boldsymbol{s}(m_0|m_1;t)$$

# Observations: Imbalanced learning between modalities




(a) ModelNet40

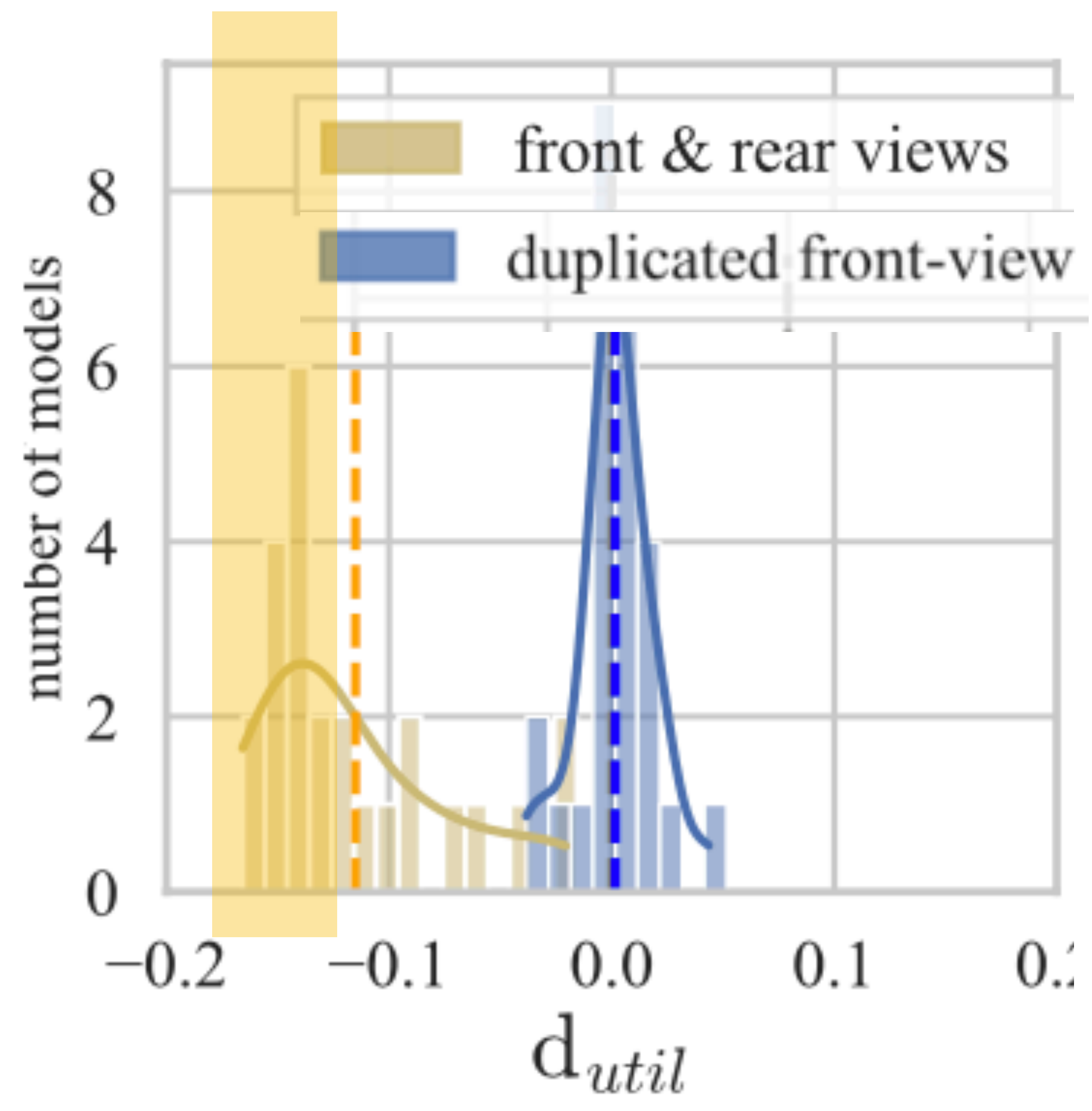



(b) NVGesture

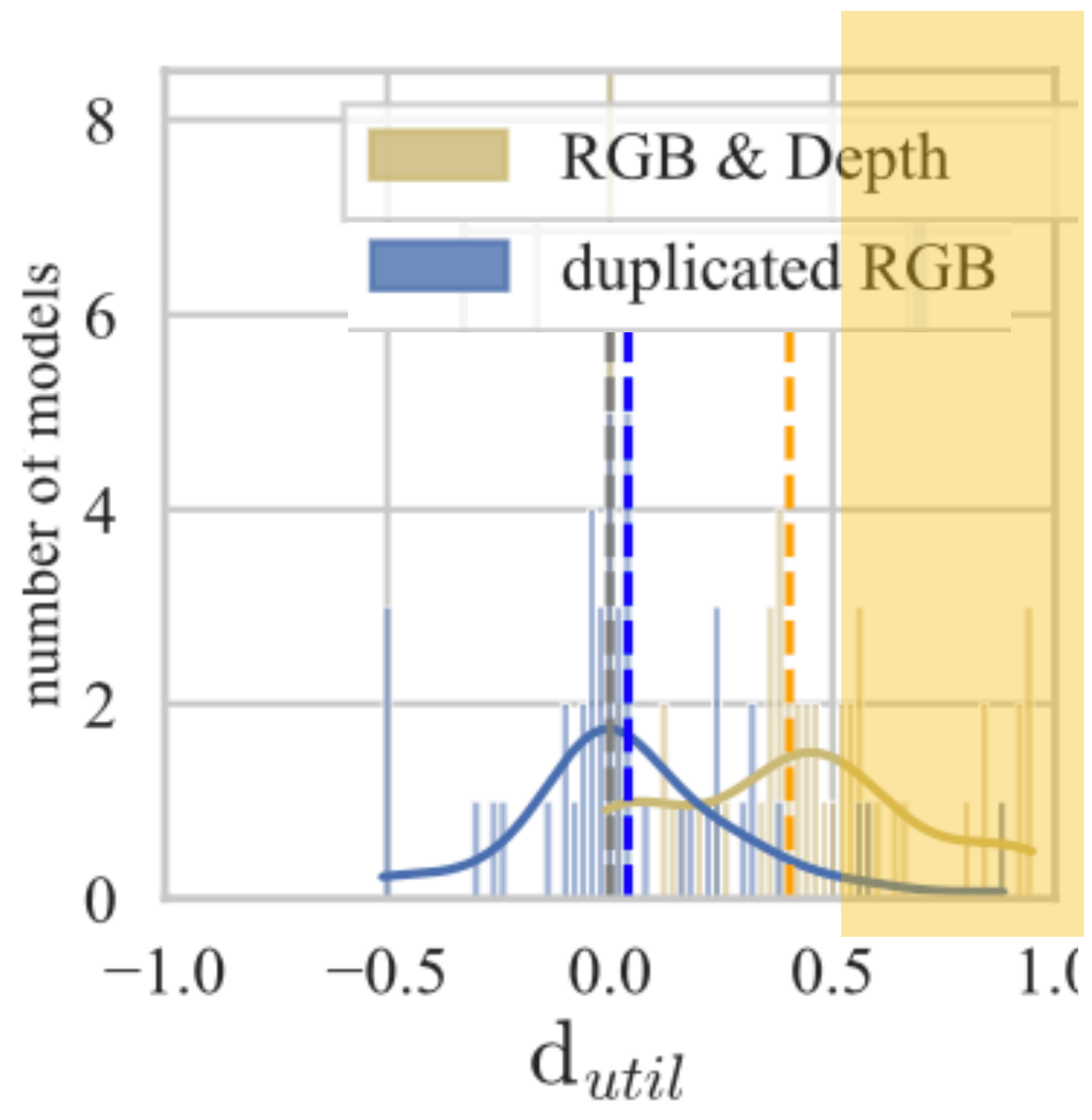
$$\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1)$$

$$\mathrm{d}_{\mathrm{speed}}(f;t) = \boldsymbol{s}(m_1|m_0;t) - \boldsymbol{s}(m_0|m_1;t)$$

# Observations: Imbalanced learning between modalities




(a) ModelNet40

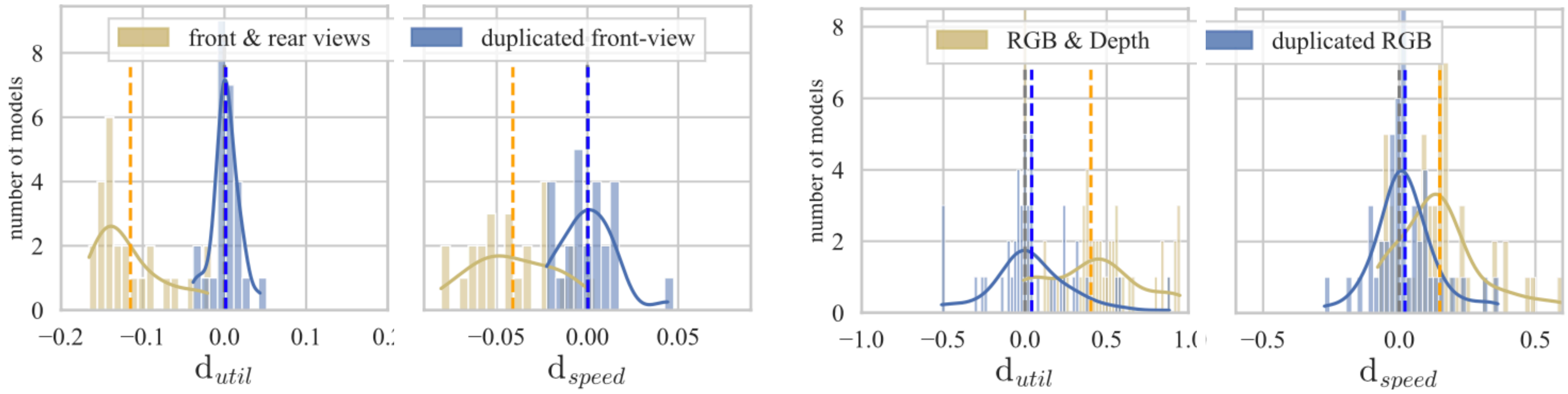



(b) NVGesture

$$\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1)$$

$$\mathrm{d}_{\mathrm{speed}}(f;t) = \boldsymbol{s}(m_1|m_0;t) - \boldsymbol{s}(m_0|m_1;t)$$

# Observations: Imbalanced learning between modalities




(a) ModelNet40 (b) NVGesture

$$\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1)$$

$$\mathrm{d}_{\mathrm{speed}}(f;t) = \boldsymbol{s}(m_1|m_0;t) - \boldsymbol{s}(m_0|m_1;t)$$

# Greedy learner hypothesis

**A multi-modal learning process is greedy when it produces models that rely on only one of the available modalities.**

# Greedy learner hypothesis

**A multi-modal learning process is greedy when it produces models that rely on only one of the available modalities.**

**The modality that the multi-modal DNN primarily relies on is the modality that is the fastest to learn from.**

# Greedy learner hypothesis

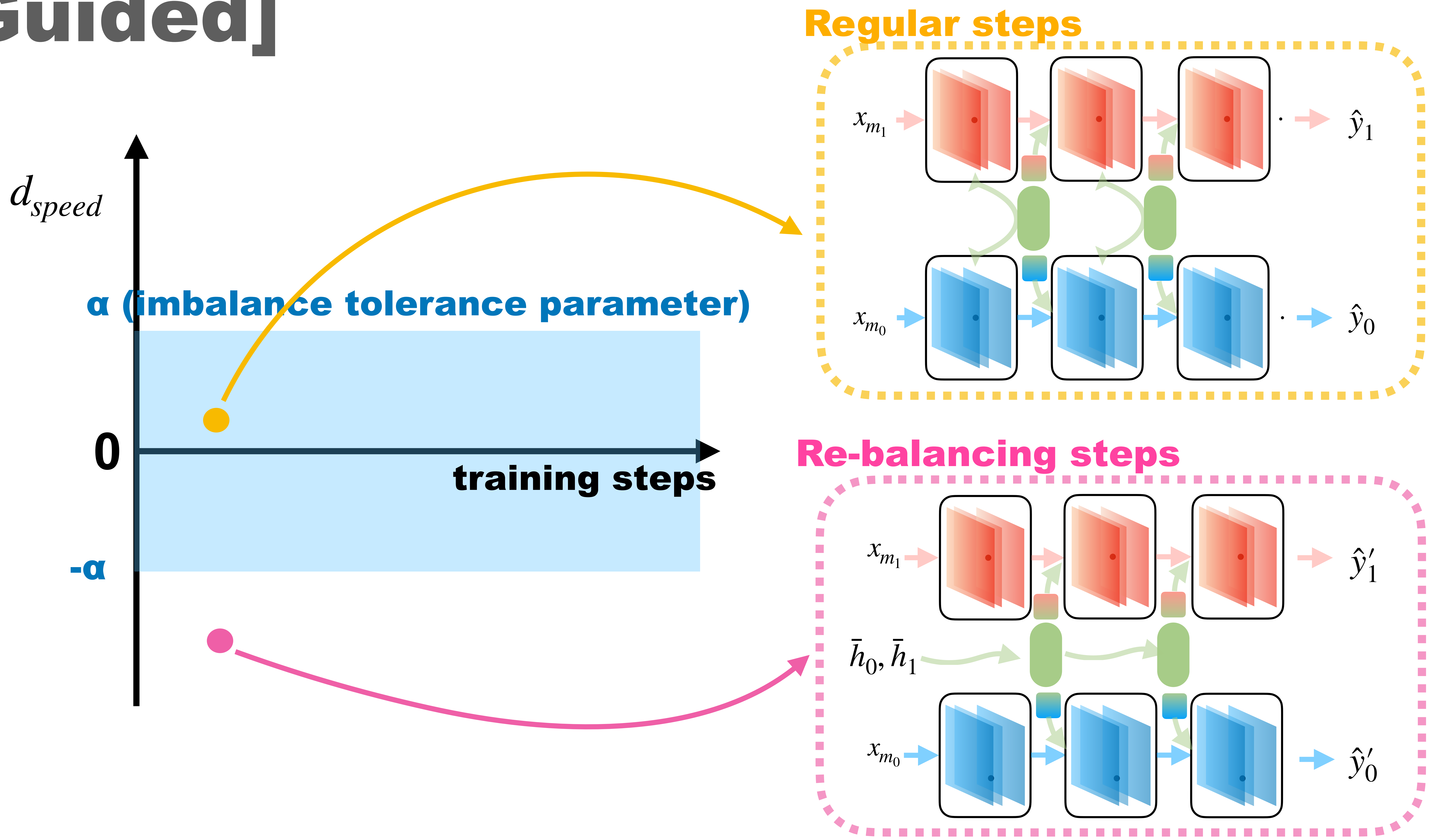
**A multi-modal learning process is greedy when it produces models that rely on only one of the available modalities.**

**The modality that the multi-modal DNN primarily relies on is the modality that is the fastest to learn from.**

**We hypothesize that a multi-modal learning process, in which a multi-modal DNN is trained to minimize the sum of the modality-specific losses, is greedy.**

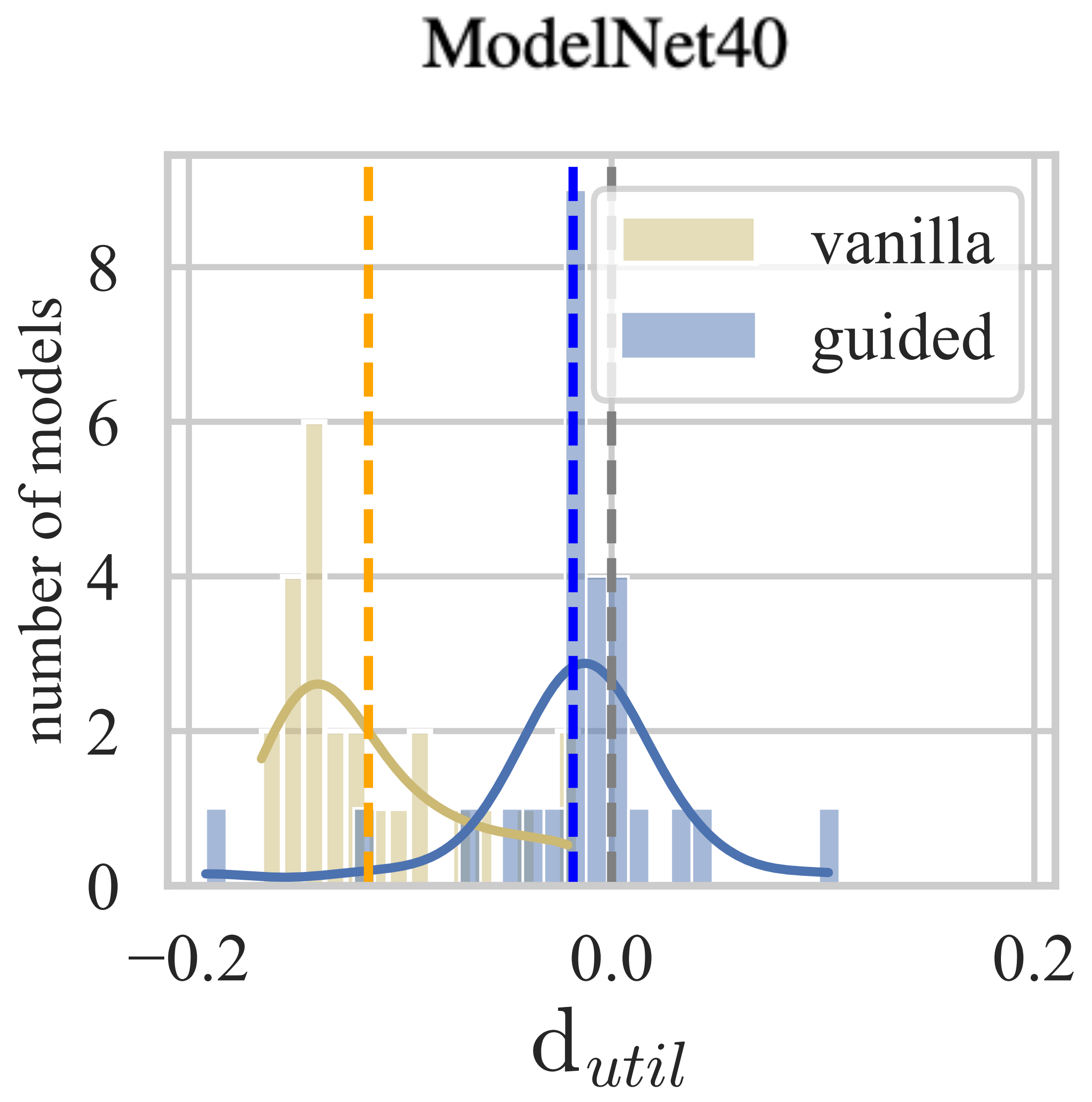
# Balanced multimodal learning

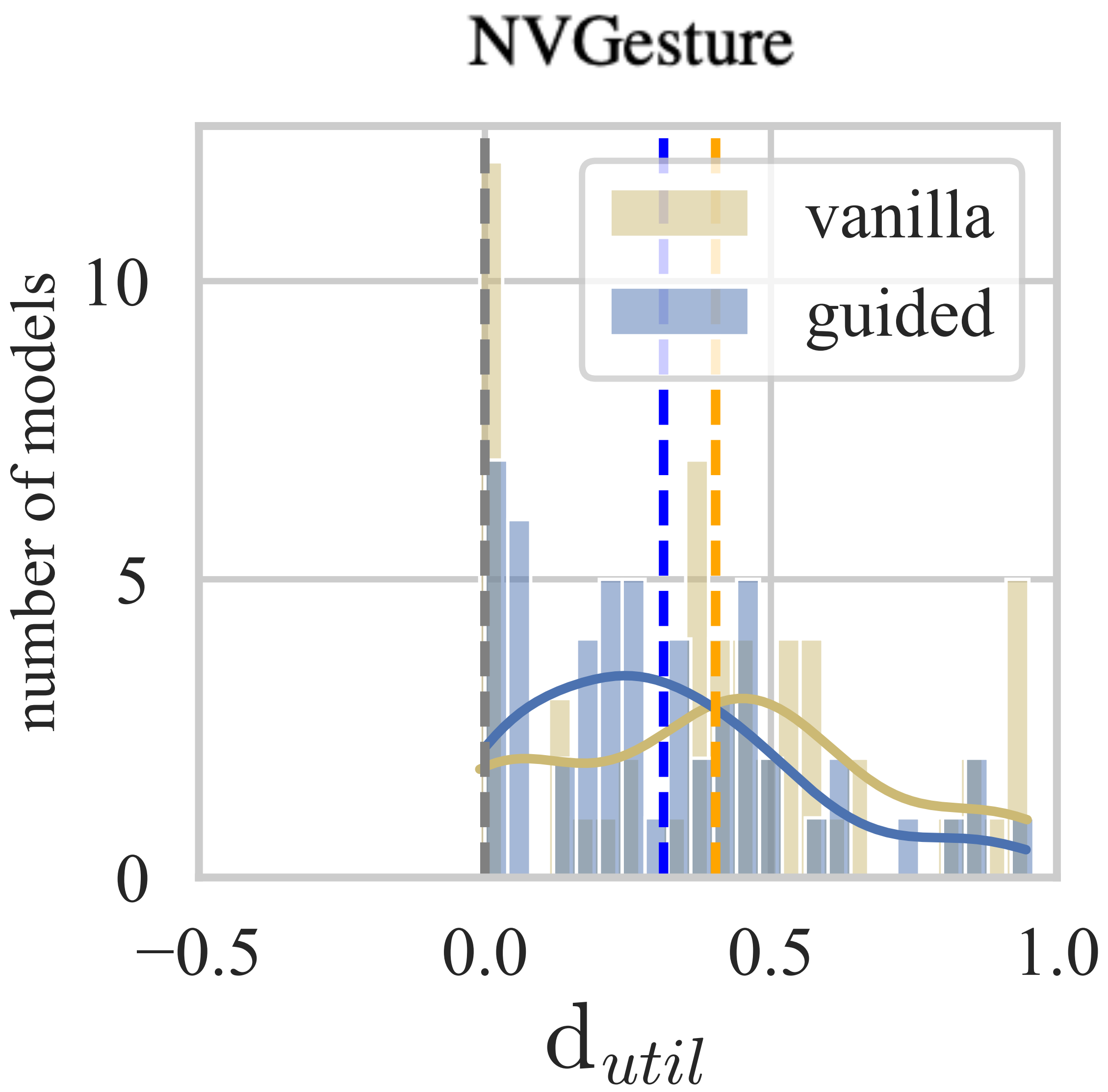
## [Guided]



# Results

- **Calibrating modality utilization**




- **Improving generalization**

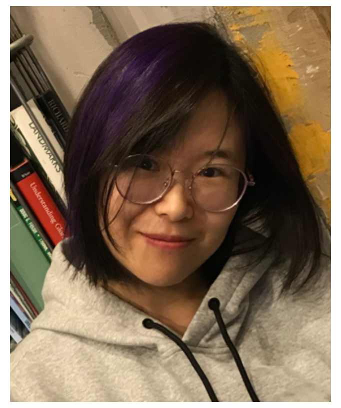
| | ModelNet40 | NVGesture-scratch | NVGesture-pretrained |
|---|---|---|---|
| uni-modal (best) | 89.34±0.39 | 77.59±0.55 | 78.98±2.02 |
| multi-modal (vanilla) | 90.09±0.58 | 79.81±1.14 | 83.20±0.21 |
| + RUBi (Cadene et al., 2019) | 90.45±0.58 | 79.95±0.12 | 81.60±1.28 |
| + random (proposed) | 91.36±0.10 | 79.88±0.90 | 82.64±0.84 |
| + guided (proposed) | **91.37±0.28** | **80.22±0.73** | **83.82±1.45** |

# Thank you!

**Nan Wu** [email: **nan.wu@nyu.edu**;   twitter:  **Nan Wu** @NanWu_ ],
**Stanisław Jastrzębski, Kyunghyun Cho and Krzysztof J. Geras.**
**Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks.** **ICML, 2022.**

https://github.com/nyukat/greedy_multimodal_learning

NYU DATA SCIENCE




CAI²R