Contrasting with Symile

Simple model-agnostic representation learning for unlimited modalities

Adriel Saporta • Aahlad Puli • Mark Goldstein • Rajesh Ranganath

New York University

representation learning with CLIP

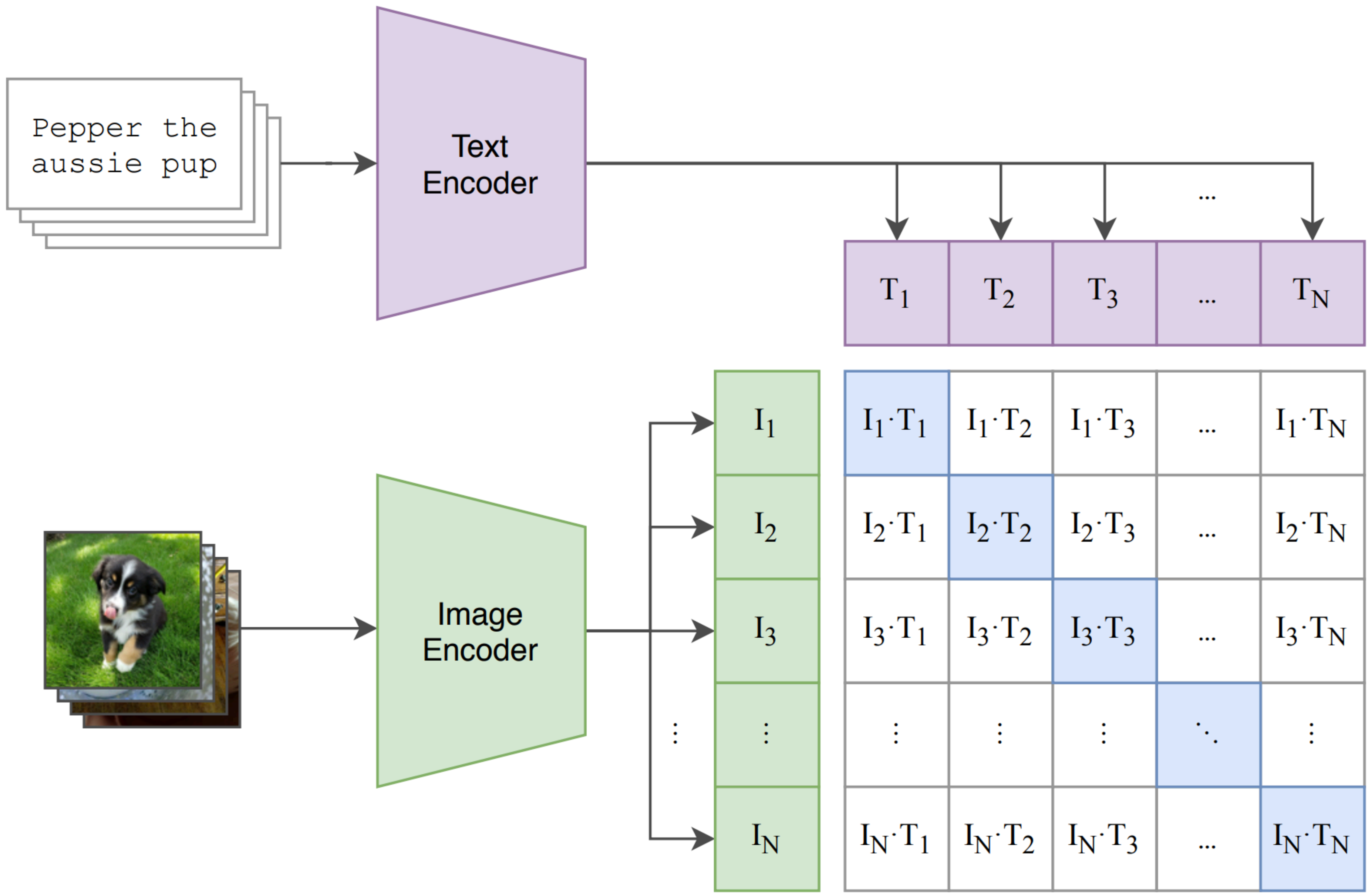
contrastive learning with InfoNCE…

$$\ell^{(\mathbf{x}\to\mathbf{y})}(\boldsymbol{\theta},\tau) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left[\left(f_{\mathbf{x}}^{\boldsymbol{\theta}}(\mathbf{x}_i)^\top f_{\mathbf{y}}^{\boldsymbol{\theta}}(\mathbf{y}_i)\right)/\tau\right]}{\sum_{j=1}^{N}\exp\left[\left(f_{\mathbf{x}}^{\boldsymbol{\theta}}(\mathbf{x}_i)^\top f_{\mathbf{y}}^{\boldsymbol{\theta}}(\mathbf{y}_j)\right)/\tau\right]}$$

$$\mathcal{L}_{\text{CLIP}}^{(\mathbf{x},\mathbf{y})}(\boldsymbol{\theta},\tau) = \frac{1}{2}\left[\ell^{(\mathbf{x}\to\mathbf{y})}(\boldsymbol{\theta},\tau) + \ell^{(\mathbf{y}\to\mathbf{x})}(\boldsymbol{\theta},\tau)\right]$$

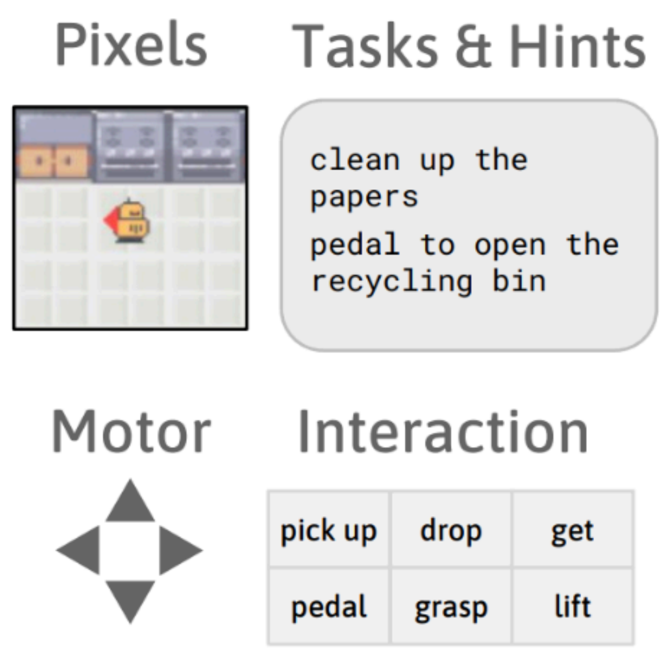
…maximizes the information between modalities

$$\mathbf{I}(\mathbf{x};\mathbf{y}) \geq \ell^{(\mathbf{x}\to\mathbf{y})}(\boldsymbol{\theta},\tau)$$

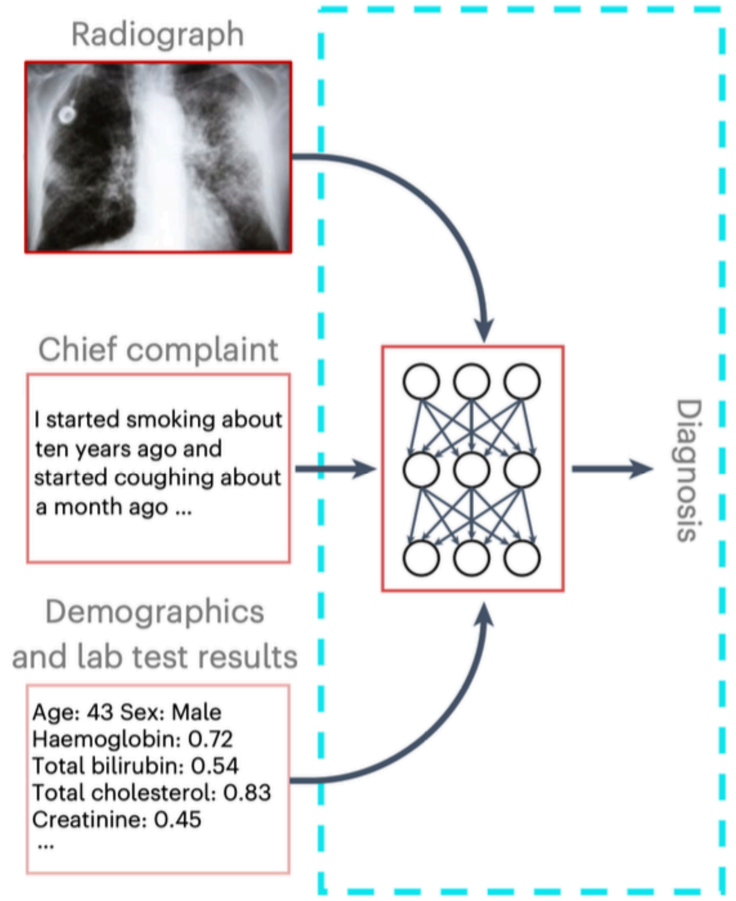


what if you have more than two modalities?

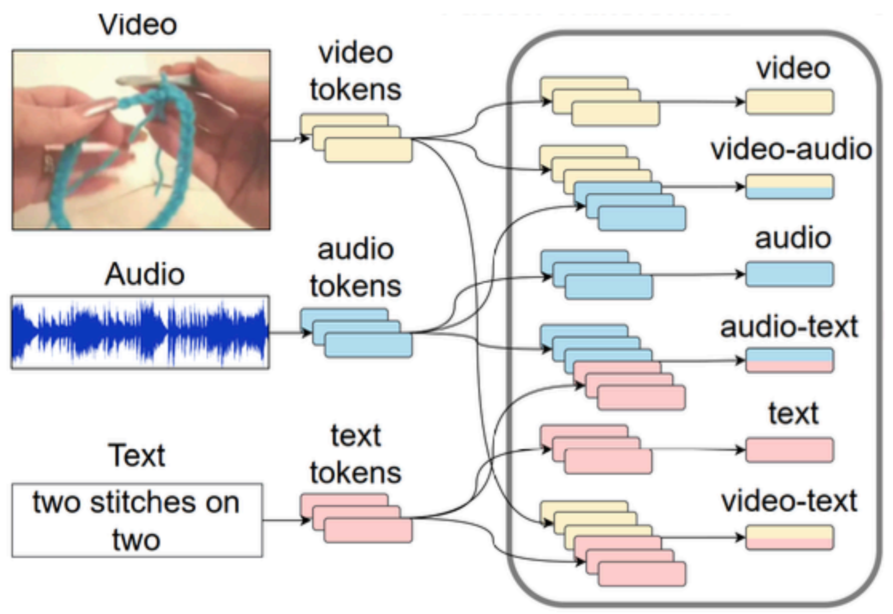
robotics



healthcare



video

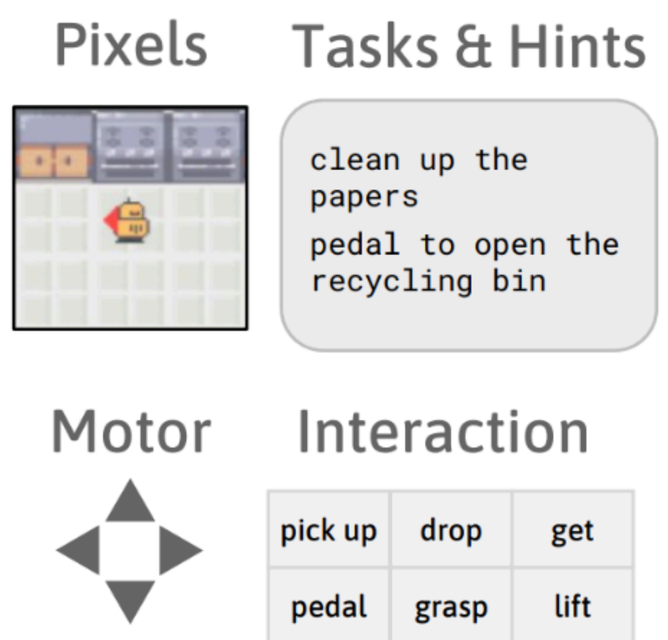


multimodal fusion models?

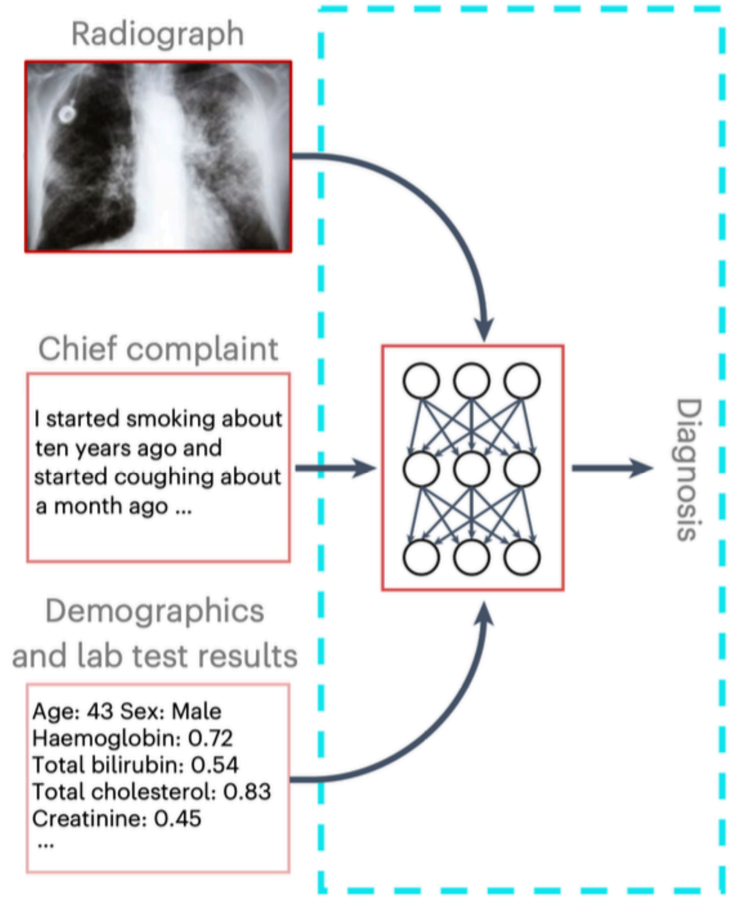
requires specialized architectures, increases operational complexity, loses modality-specific representations

what if you have more than two modalities?

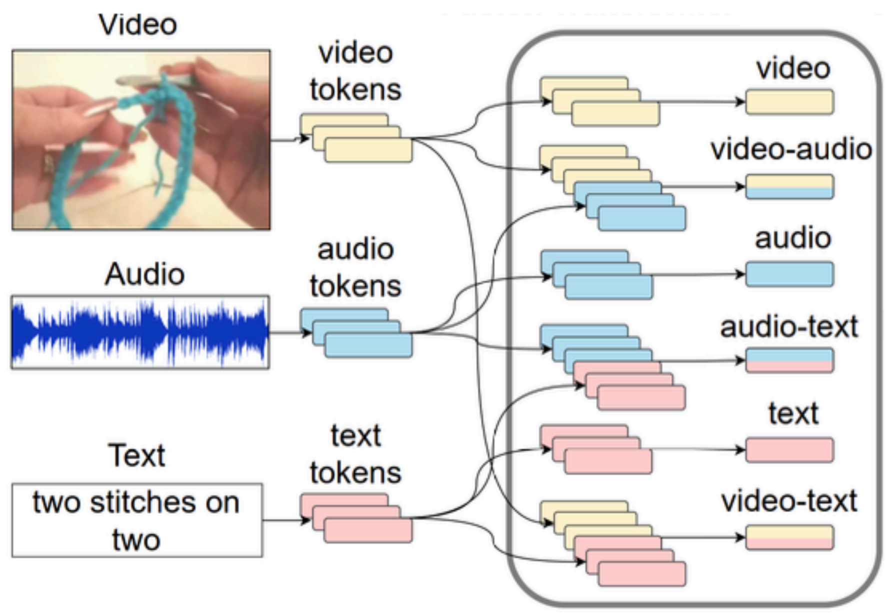
robotics



healthcare



video

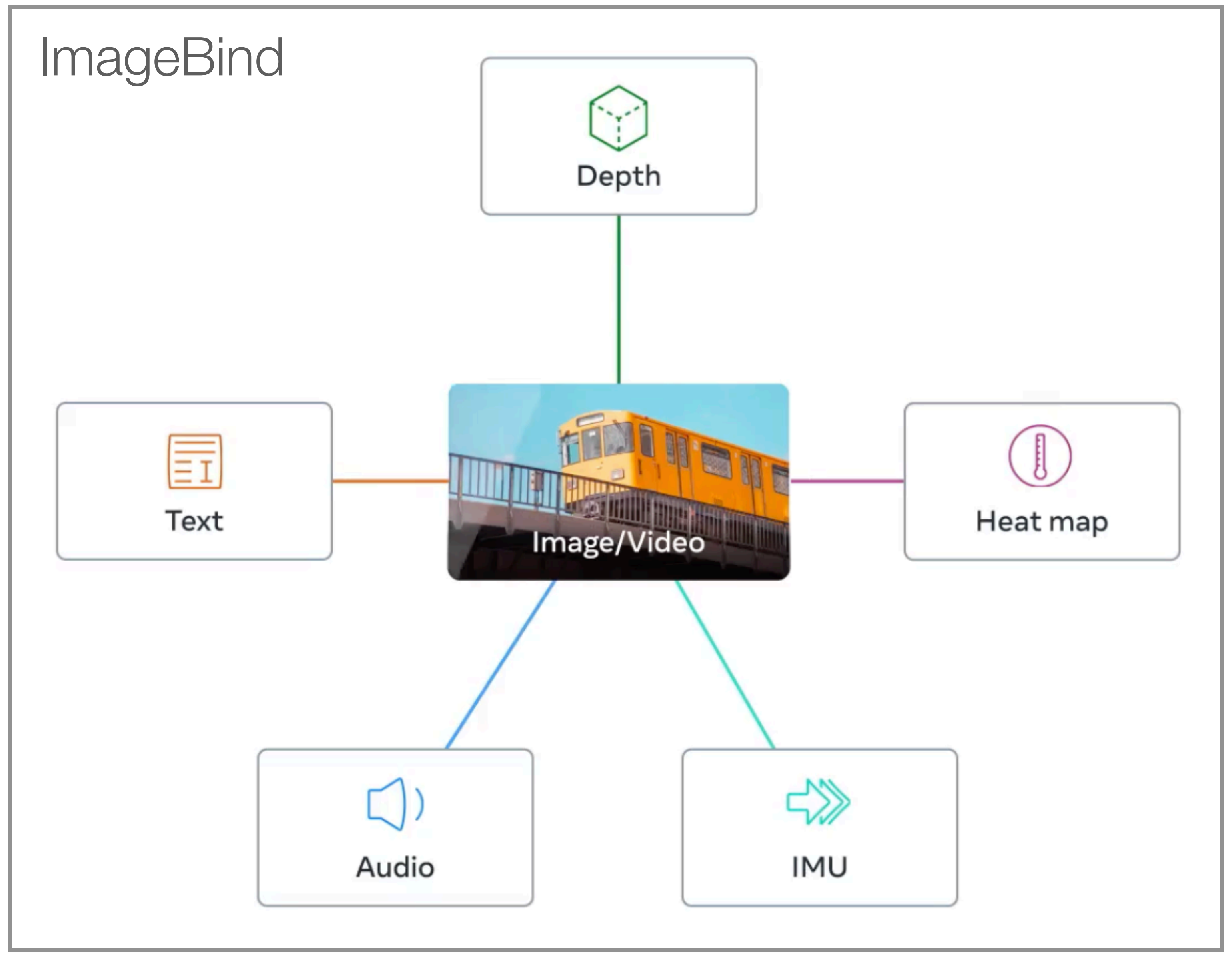


multimodal fusion models?

requires specialized architectures, increases operational complexity, loses modality-specific representations

pairwise CLIP?

$$\mathcal{L}_{\text{CLIP}}^{(\mathbf{x},\mathbf{y},\mathbf{z})}(\boldsymbol{\theta},\tau) = \mathcal{L}_{\text{CLIP}}^{(\mathbf{x},\mathbf{y})}(\boldsymbol{\theta},\tau) + \mathcal{L}_{\text{CLIP}}^{(\mathbf{y},\mathbf{z})}(\boldsymbol{\theta},\tau) + \mathcal{L}_{\text{CLIP}}^{(\mathbf{x},\mathbf{z})}(\boldsymbol{\theta},\tau)$$



…could this work?

a simple task for pairwise CLIP

$\mathbf{a}, \mathbf{b} \sim \text{Bernoulli}(0.5), \quad \mathbf{c} = \mathbf{a} \text{ XOR } \mathbf{b}$

The task is to find the $\mathbf{b}$ that corresponds to a given $\mathbf{a}$ and $\mathbf{c}$. → CLIP performs no better than random chance!

$\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ are jointly *dependent*… $\mathbf{I}(\mathbf{a}; \mathbf{b} \,|\, \mathbf{c}) > 0$

…but pairwise *independent* $\mathbf{I}(\mathbf{a}; \mathbf{b}) = \mathbf{I}(\mathbf{b}; \mathbf{c}) = \mathbf{I}(\mathbf{a}; \mathbf{c}) = 0$ → There's nothing for CLIP to learn!

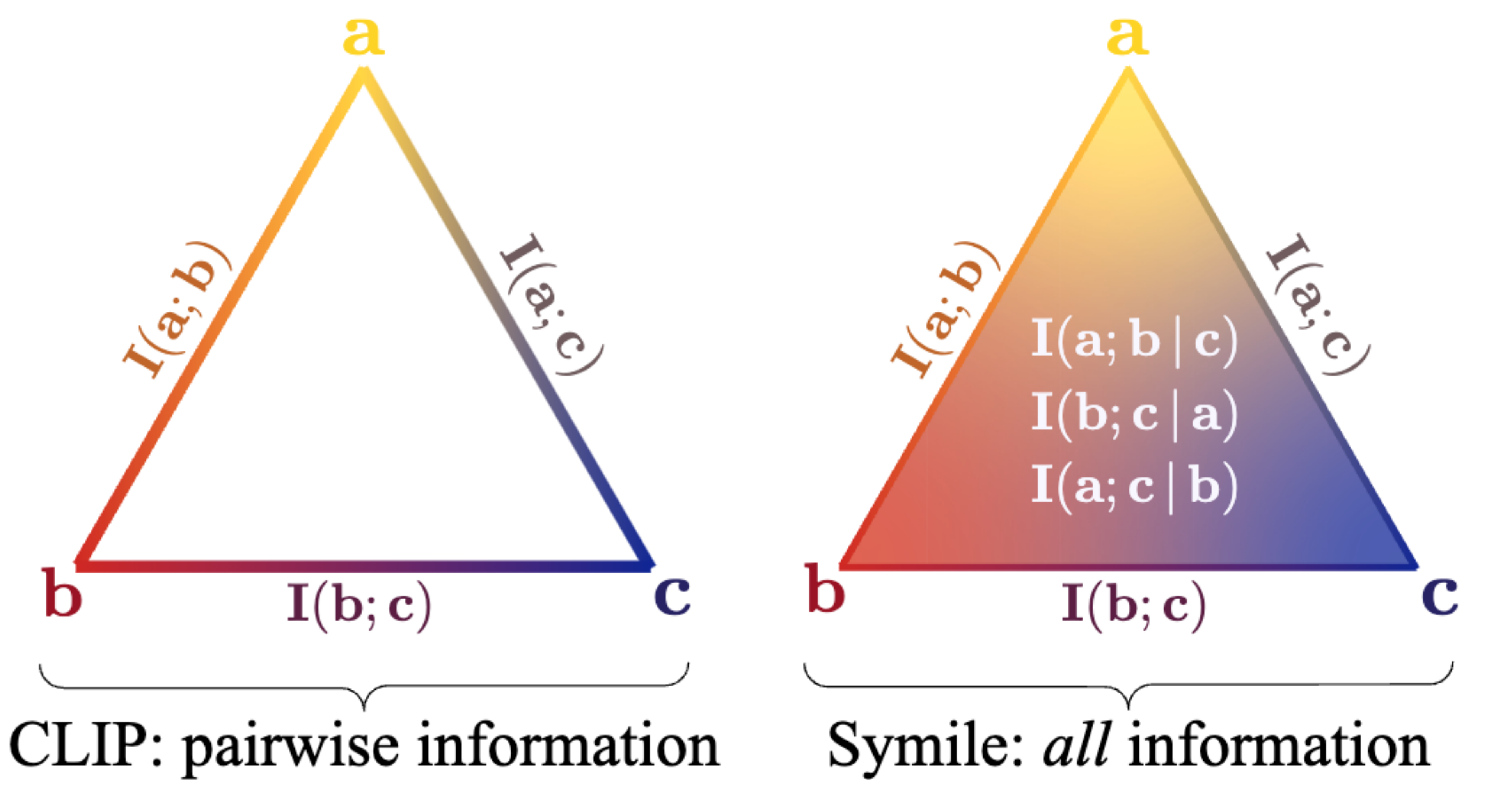
For multimodal representation learning, we need an objective that…

- is as simple as CLIP
- learns architecture-agnostic and modality-specific representations
- captures higher-order information between any number of modalities

Symile targets total correlation

$$\mathbf{TC}(\mathbf{x}_1, \ldots, \mathbf{x}_M) = D_{\mathrm{KL}}\big(p(\mathbf{x}_1, \ldots, \mathbf{x}_M) \,\|\, p(\mathbf{x}_1) \cdots p(\mathbf{x}_M)\big)$$

$$3 \cdot \underbrace{\mathbf{TC}(\mathbf{x}, \mathbf{y}, \mathbf{z})}_{\text{Symile target}} = 2 \cdot \underbrace{\big[\mathbf{I}(\mathbf{x}; \mathbf{y}) + \mathbf{I}(\mathbf{y}; \mathbf{z}) + \mathbf{I}(\mathbf{x}; \mathbf{z})\big]}_{\substack{\text{pairwise information} \\ \text{(CLIP target)}}} + \underbrace{\mathbf{I}(\mathbf{x}; \mathbf{y} \,|\, \mathbf{z}) + \mathbf{I}(\mathbf{y}; \mathbf{z} \,|\, \mathbf{x}) + \mathbf{I}(\mathbf{x}; \mathbf{z} \,|\, \mathbf{y})}_{\text{higher-order information}}$$



Symile

$$\mathbf{TC}(\mathbf{x},\mathbf{y},\mathbf{z}) \geq \log N + \mathop{\mathbb{E}}_{p(\mathbf{x},\mathbf{Y}_N,\mathbf{Z}_N \mid \mathbf{i}=i)} \log \frac{\exp g(\mathbf{x},\mathbf{y}_i,\mathbf{z}_i)}{\sum_{j=1}^{N} \exp g(\mathbf{x},\mathbf{y}_j,\mathbf{z}_j)}$$

$$\ell^{(\mathbf{x}\rightarrow\mathbf{y},\mathbf{z})}(\boldsymbol{\theta},\tau) = -\frac{1}{N'}\sum_{i=1}^{N'} \log \frac{\exp\left(\langle f_{\mathbf{x}}^{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\mathbf{y}}^{\boldsymbol{\theta}}(\mathbf{y}_i), f_{\mathbf{z}}^{\boldsymbol{\theta}}(\mathbf{z}_i)\rangle/\tau\right)}{\exp\left(\langle f_{\mathbf{x}}^{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\mathbf{y}}^{\boldsymbol{\theta}}(\mathbf{y}_i), f_{\mathbf{z}}^{\boldsymbol{\theta}}(\mathbf{z}_i)\rangle/\tau\right) + \sum_{j=1}^{N-1} \exp\left(\langle f_{\mathbf{x}}^{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\mathbf{y}}^{\boldsymbol{\theta}}(\mathbf{y}'_j), f_{\mathbf{z}}^{\boldsymbol{\theta}}(\mathbf{z}'_j)\rangle/\tau\right)}$$

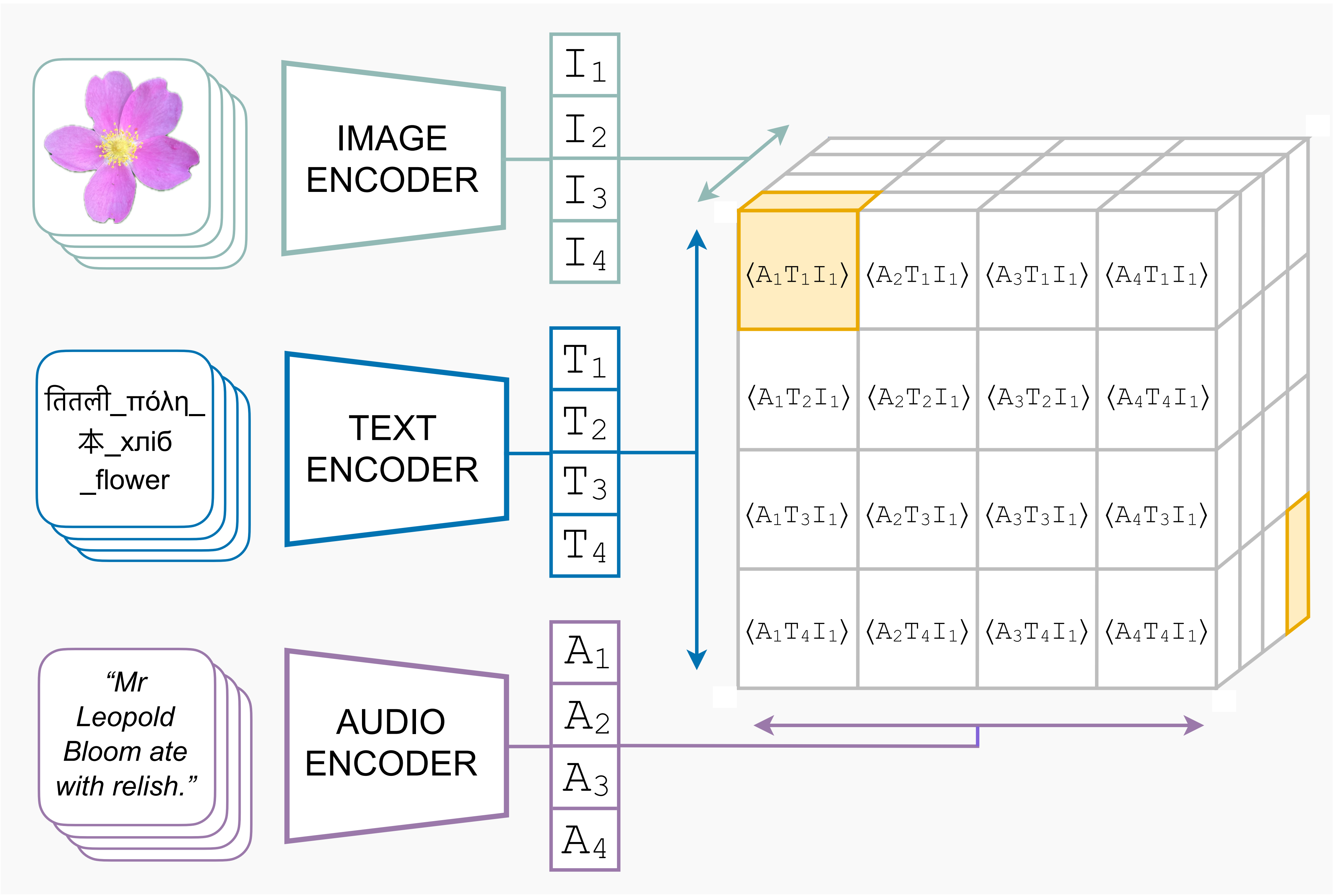
$$\mathcal{L}_{\text{Symile}}^{(\mathbf{x},\mathbf{y},\mathbf{z})}(\boldsymbol{\theta},\tau) = \frac{1}{3}\left[\ell^{(\mathbf{x}\rightarrow\mathbf{y},\mathbf{z})}(\boldsymbol{\theta},\tau) + \ell^{(\mathbf{y}\rightarrow\mathbf{x},\mathbf{z})}(\boldsymbol{\theta},\tau) + \ell^{(\mathbf{z}\rightarrow\mathbf{x},\mathbf{y})}(\boldsymbol{\theta},\tau)\right]$$

Multilinear inner product (MIP)

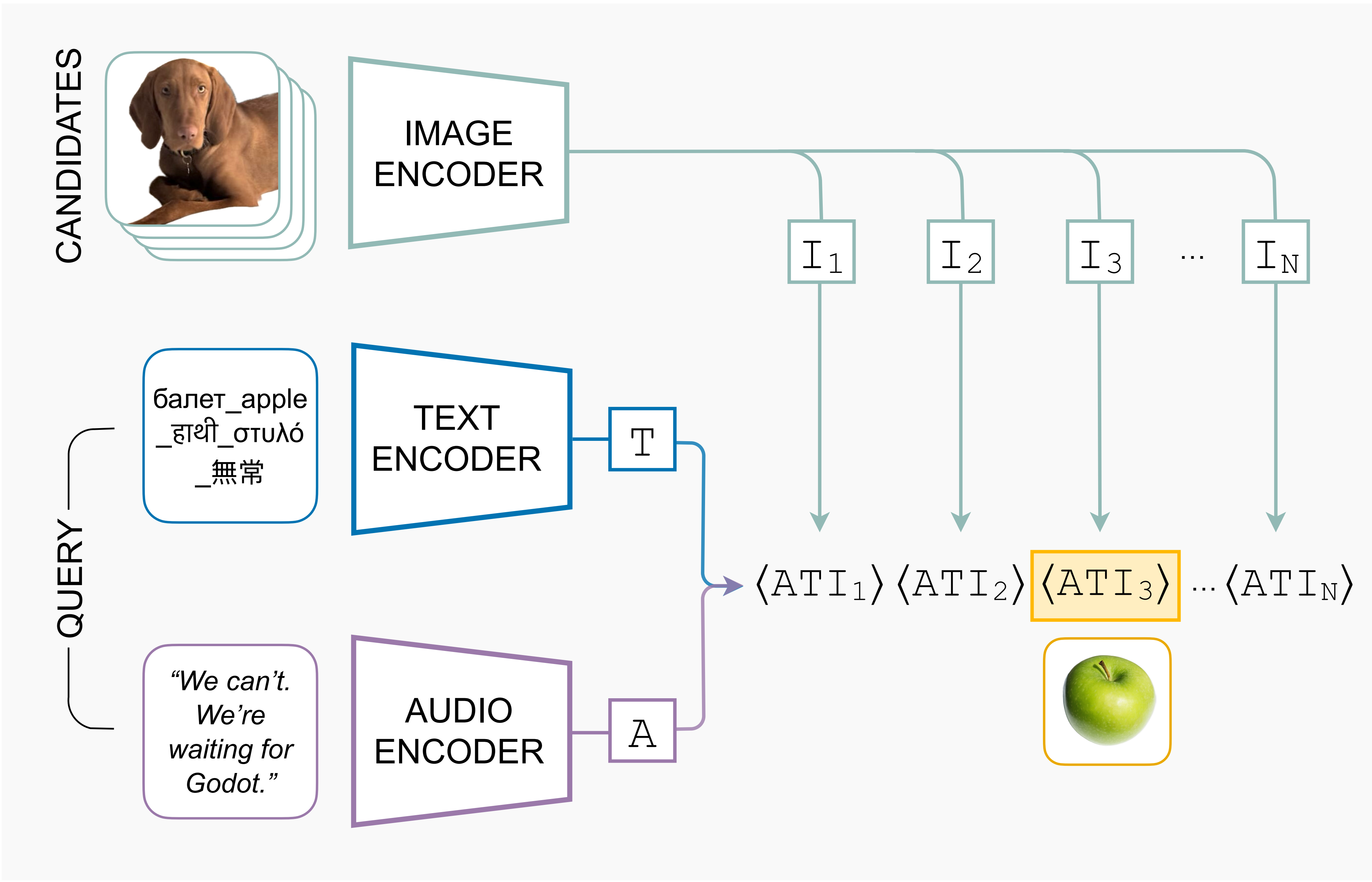
$$\langle \mathbf{xyz} \rangle = \sum_{d=1}^{D} x_d y_d z_d$$

Symile

Symile pre-training



Zero-shot prediction




```
# https://github.com/rajesh-lab/symile

pip install symile
```

```python
from symile import Symile, MIPSimilarity

symile_loss = Symile()
mip_similarity = MIPSimilarity()

# training: compute loss with embeddings a, b, c
loss = symile_loss([a, b, c], logit_scale_exp)

# evaluation: compute similarity scores for zero-shot retrieval
scores = mip_similarity(candidates_a, [query_b, query_c])
```

revisiting that XOR experiment

$$a_j, b_j \sim \text{Bernoulli}(0.5), \quad i \sim \text{Bernoulli}(\hat{p}), \quad c_j = (a_j \text{ XOR } b_j)^i \cdot a_j^{(1-i)}$$

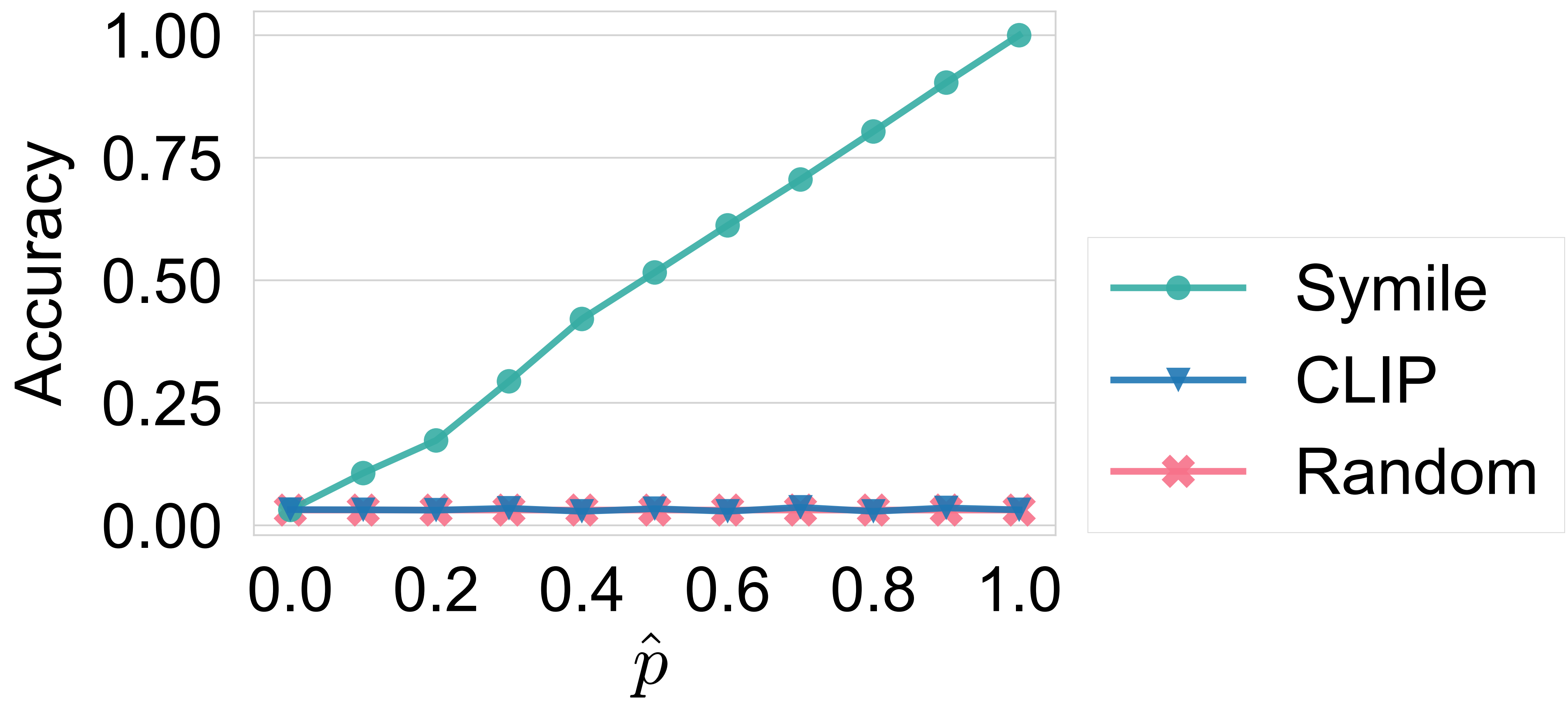
$$\mathbf{a} = [a_1, \dots, a_5], \quad \mathbf{b} = [b_1, \dots, b_5], \quad \mathbf{c} = [c_1, \dots, c_5]$$

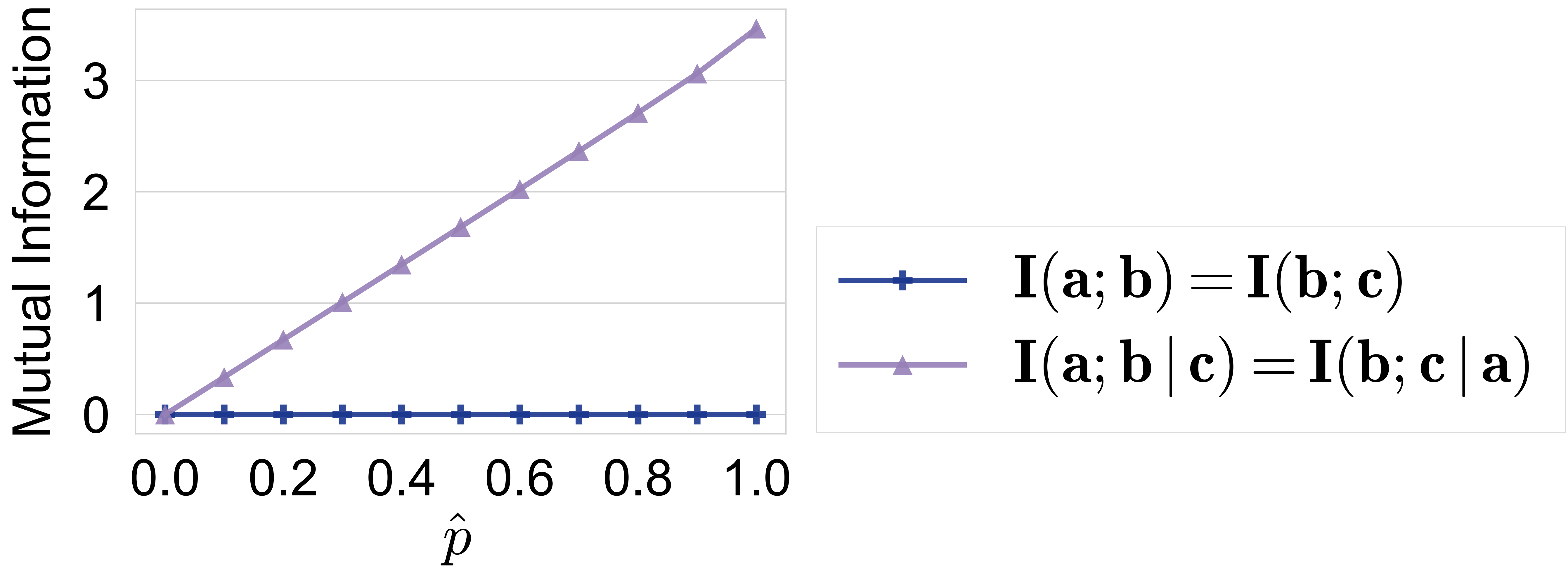
→

$$\text{when } \hat{p} = 0, \ \ c_j = a_j$$

$$\text{when } \hat{p} = 1, \ \ c_j = a_j \text{ XOR } b_j$$

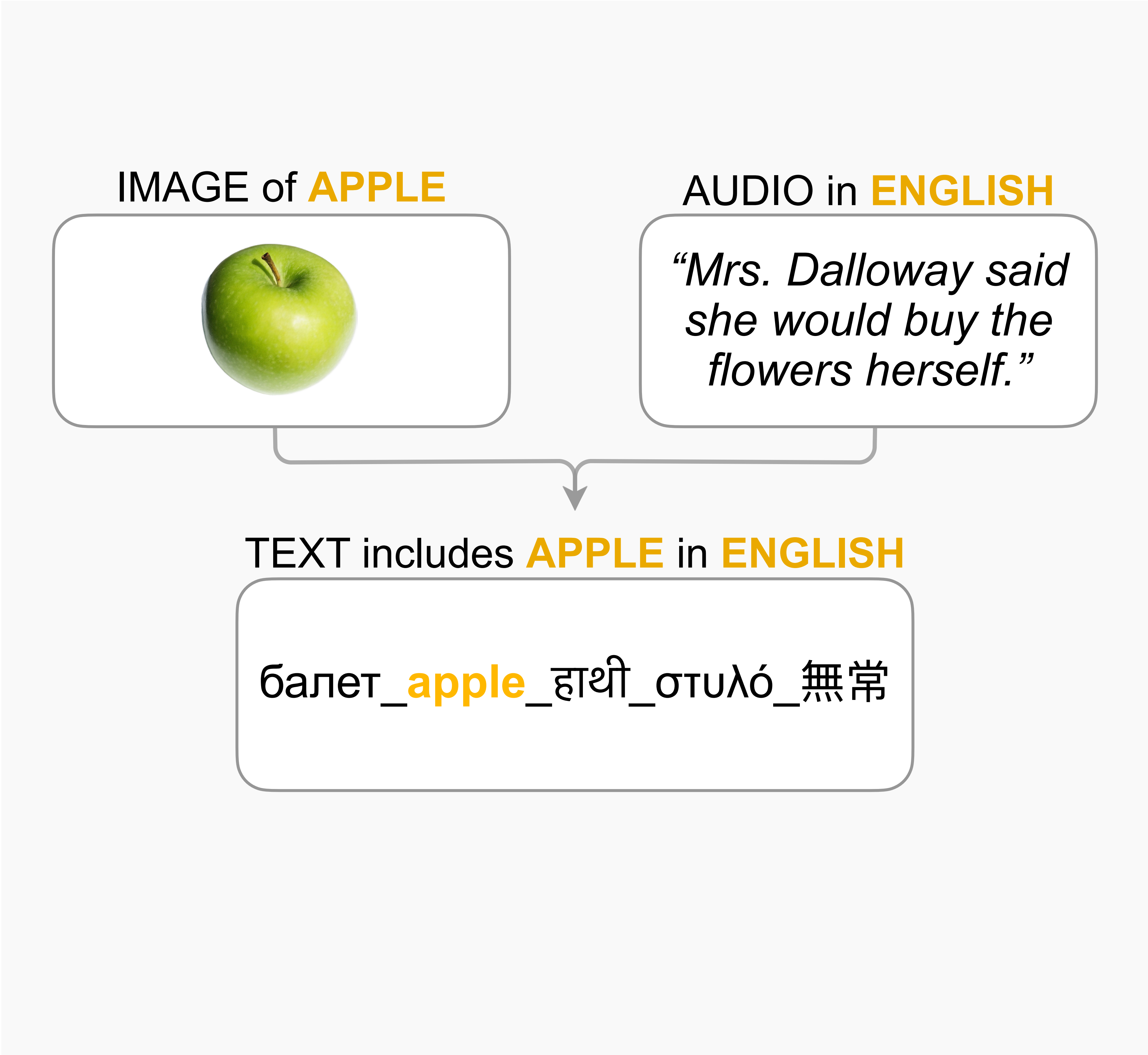
Task is to find the $\mathbf{b}$ that corresponds to a given $\mathbf{a}$ and $\mathbf{c}$.





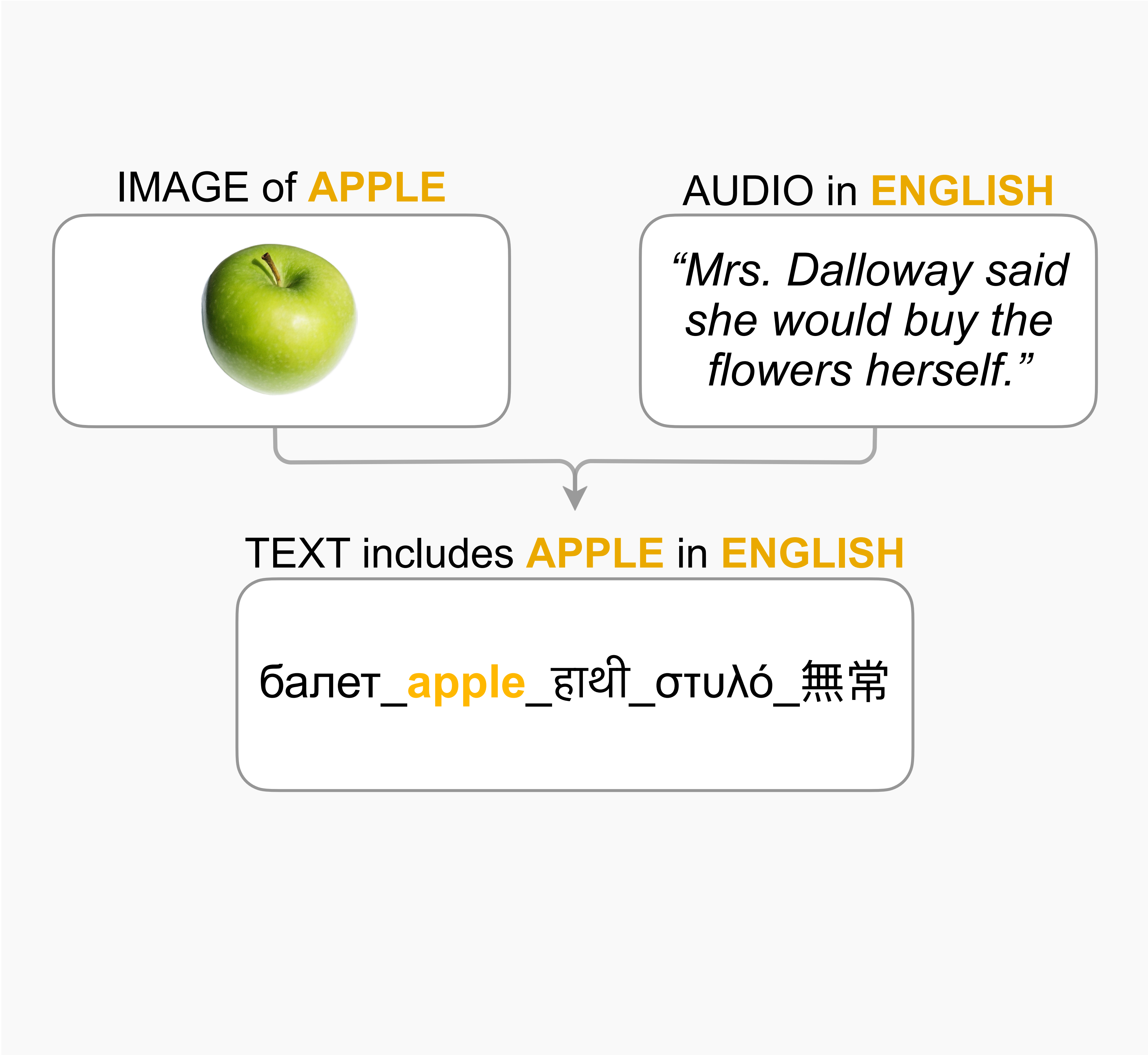
Symile-M3: a new multimodal dataset

Data generation

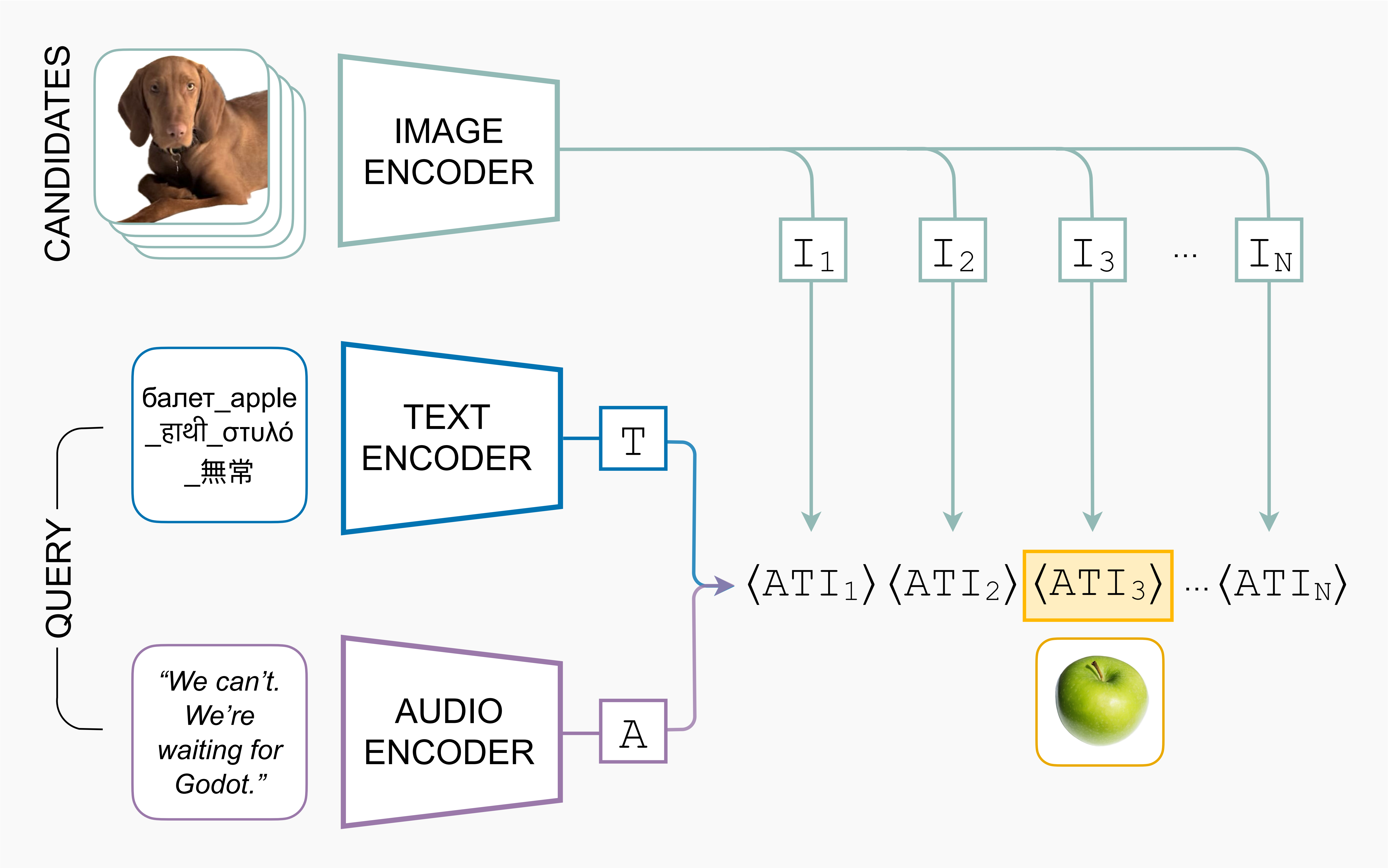


Symile-M3: a new multimodal dataset

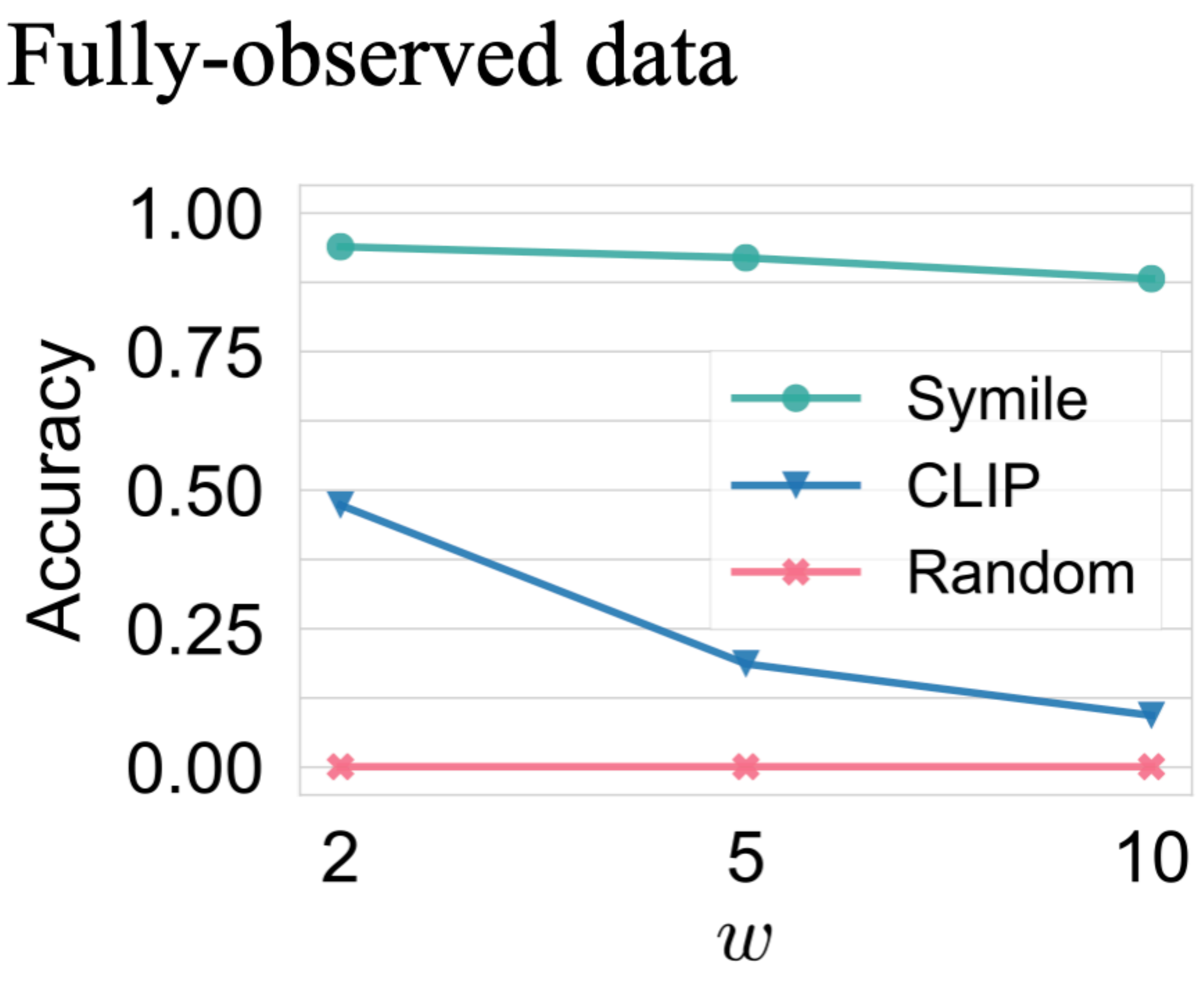
Data generation

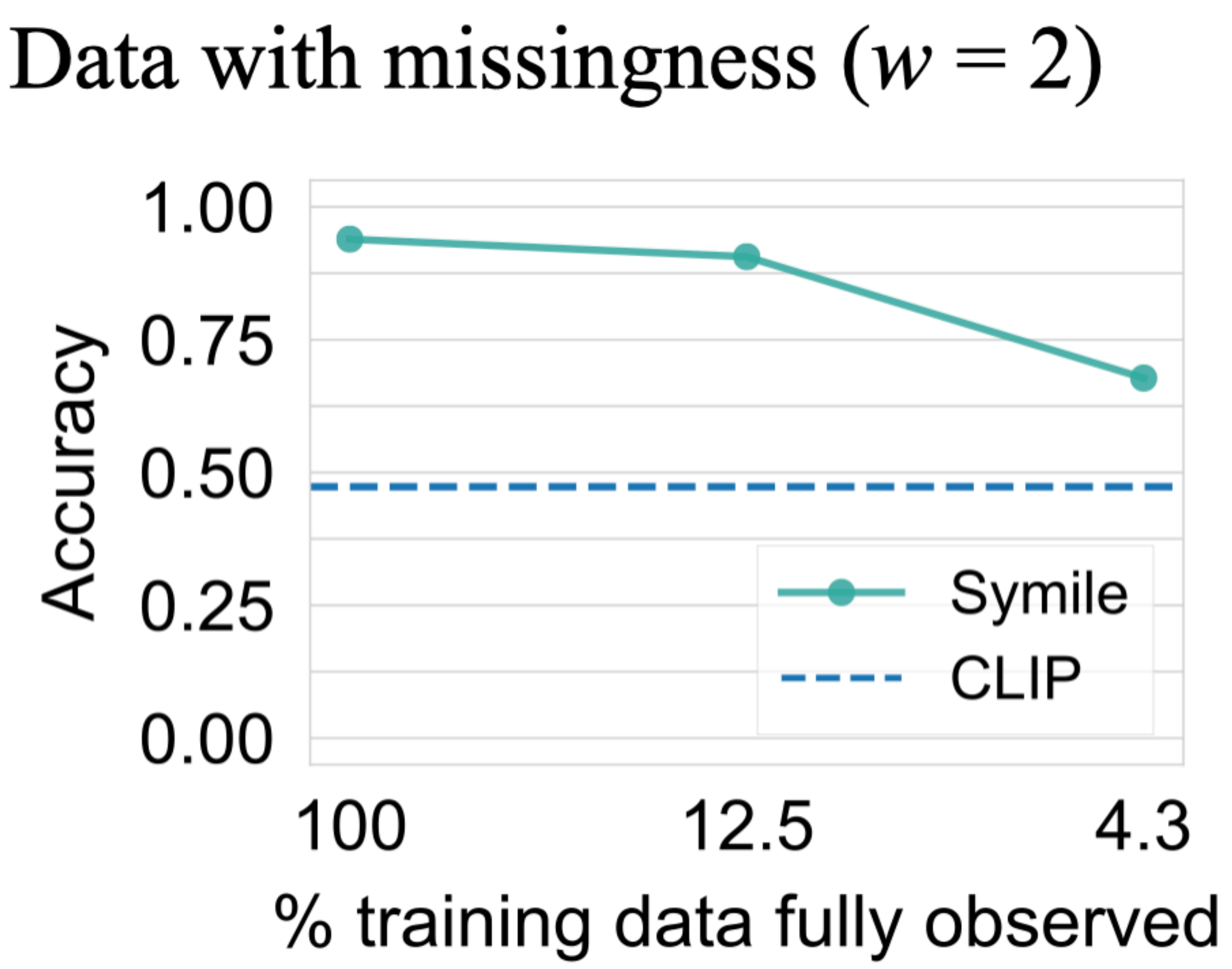


Zero-shot prediction



Symile-M3: a new multimodal dataset

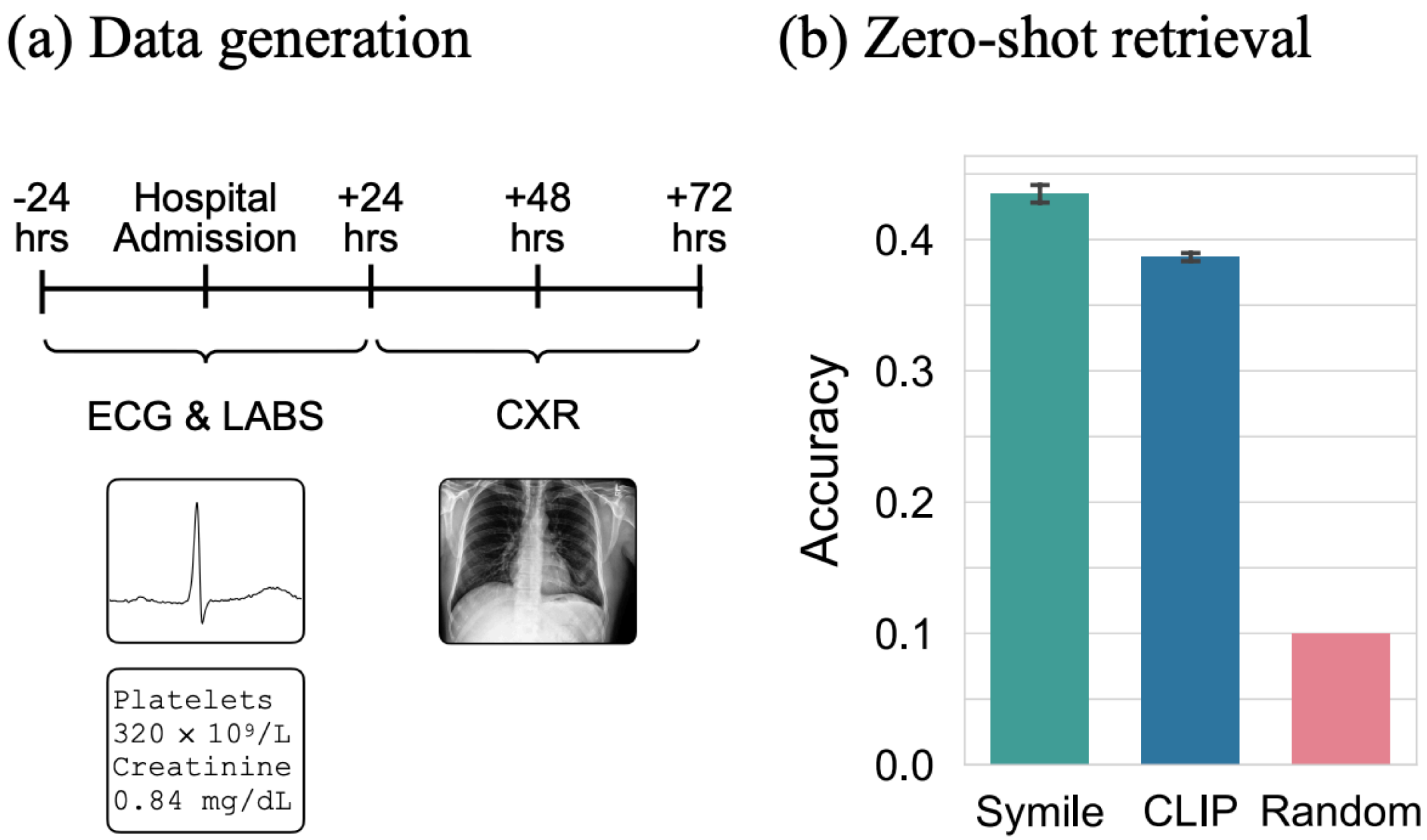




but wait, there's more!

Check out our paper for more methodological and experimental contributions:

- Efficient negative sampling strategies
- Learning sufficient statistics with Symile
- Experimental results on a new multimodal clinical dataset (see right)



https://arxiv.org/abs/2411.01053
https://github.com/rajesh-lab/symile