Vision CNNs trained to estimate spatial latents learned similar ventral-stream-aligned representations

ICLR 2025

Vision is “to know **what** is **where** by looking”

— David Marr

Image from Marr, D., 2010.

Vision is much more than object recognition

Vision is much more than object recognition

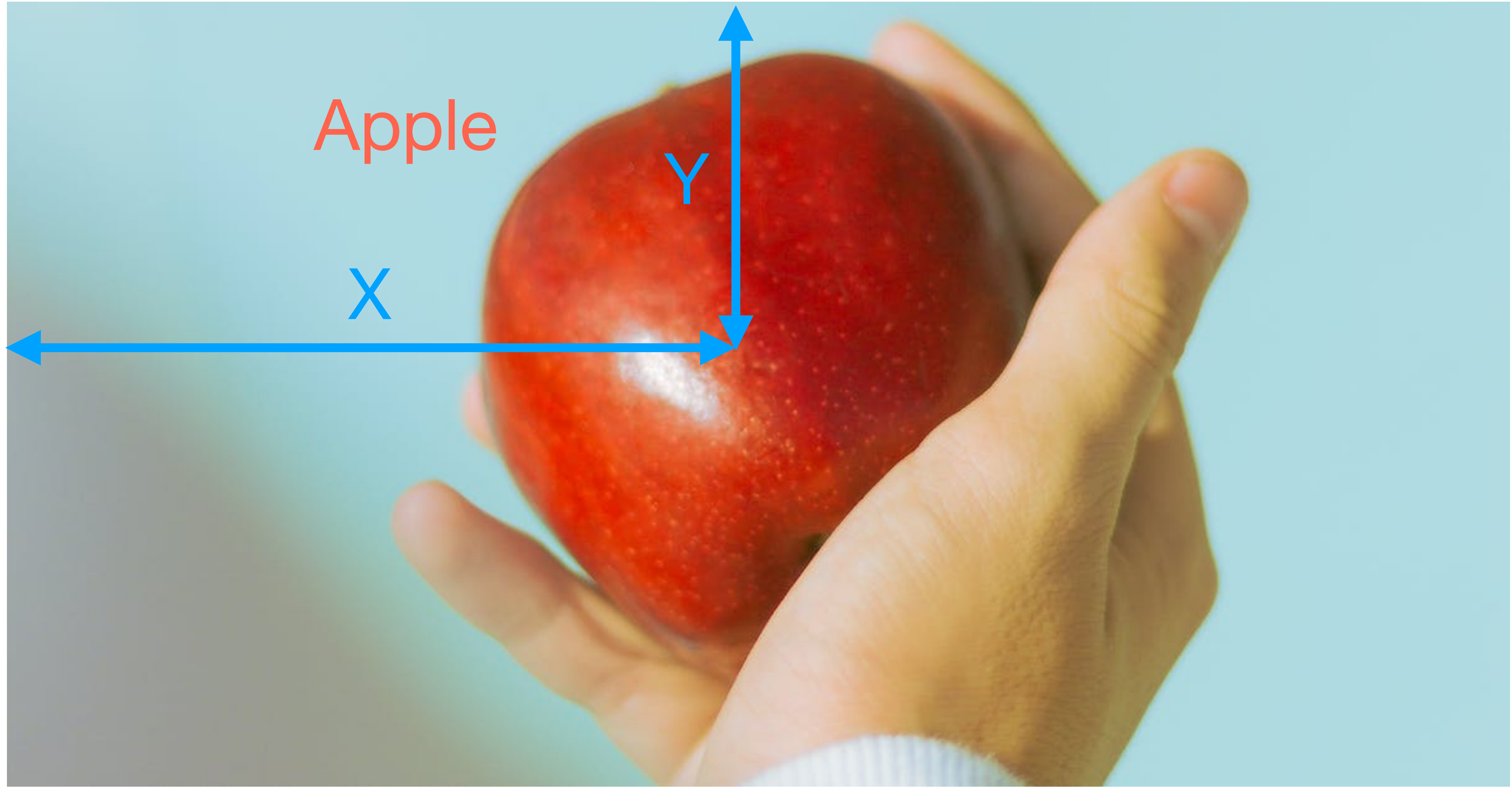
Vision is much more than object recognition

Vision is much more than object recognition

- When we see an object, we don't just see an abstract category.

Vision is much more than object recognition

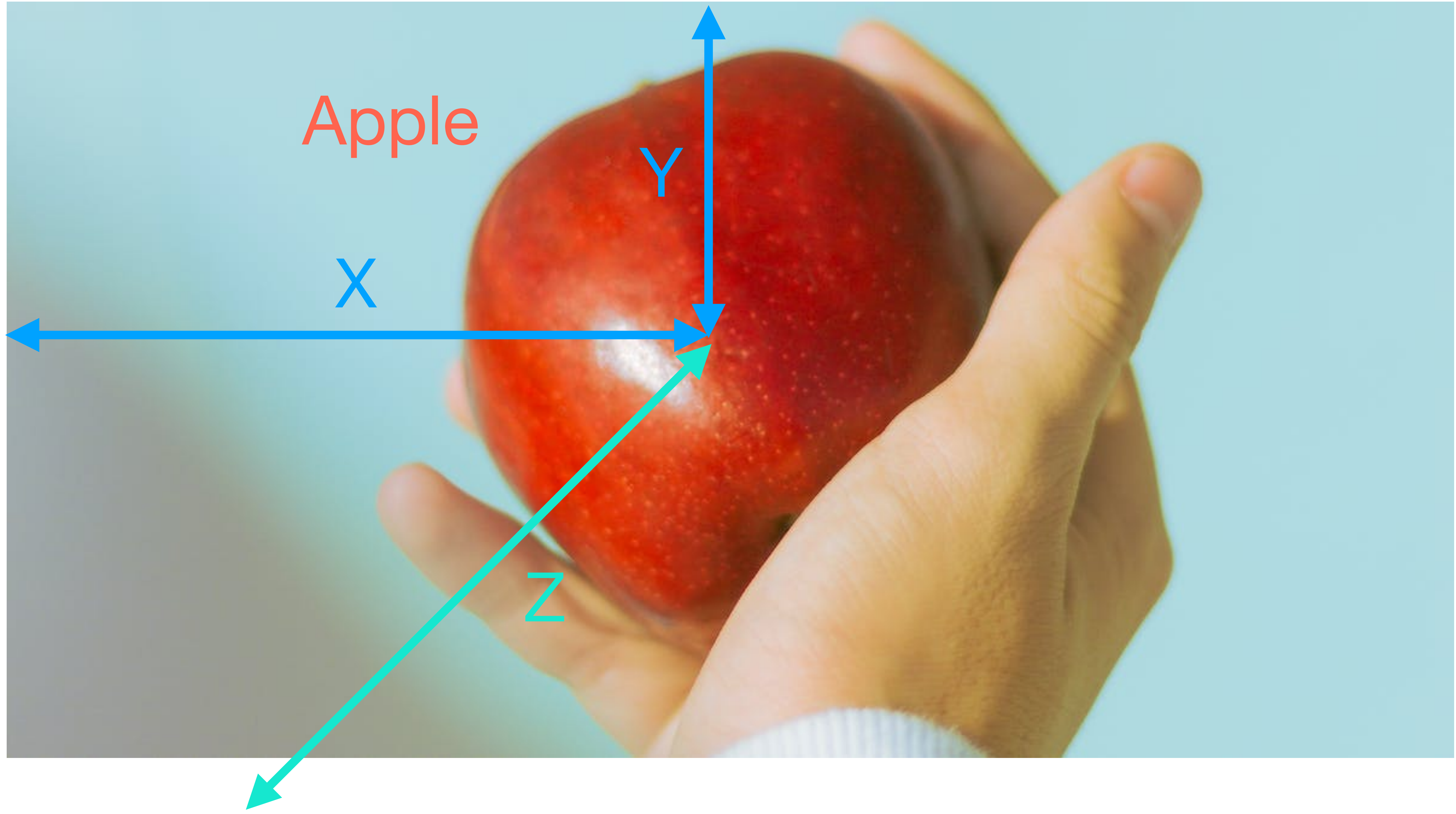
- When we see an object, we don’t just see an abstract category.



Location

Vision is much more than object recognition

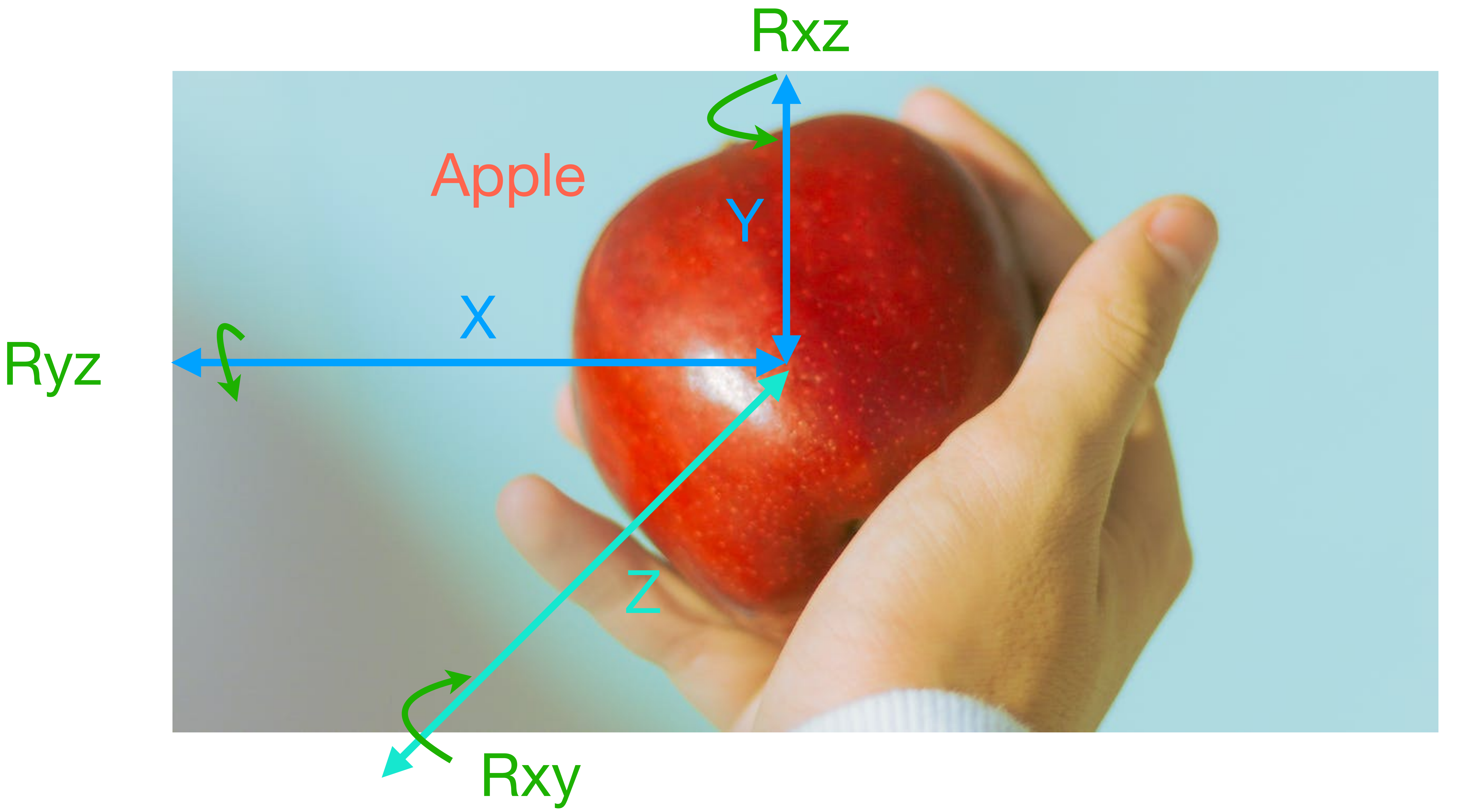
- When we see an object, we don't just see an abstract category.



Location Distance

Vision is much more than object recognition

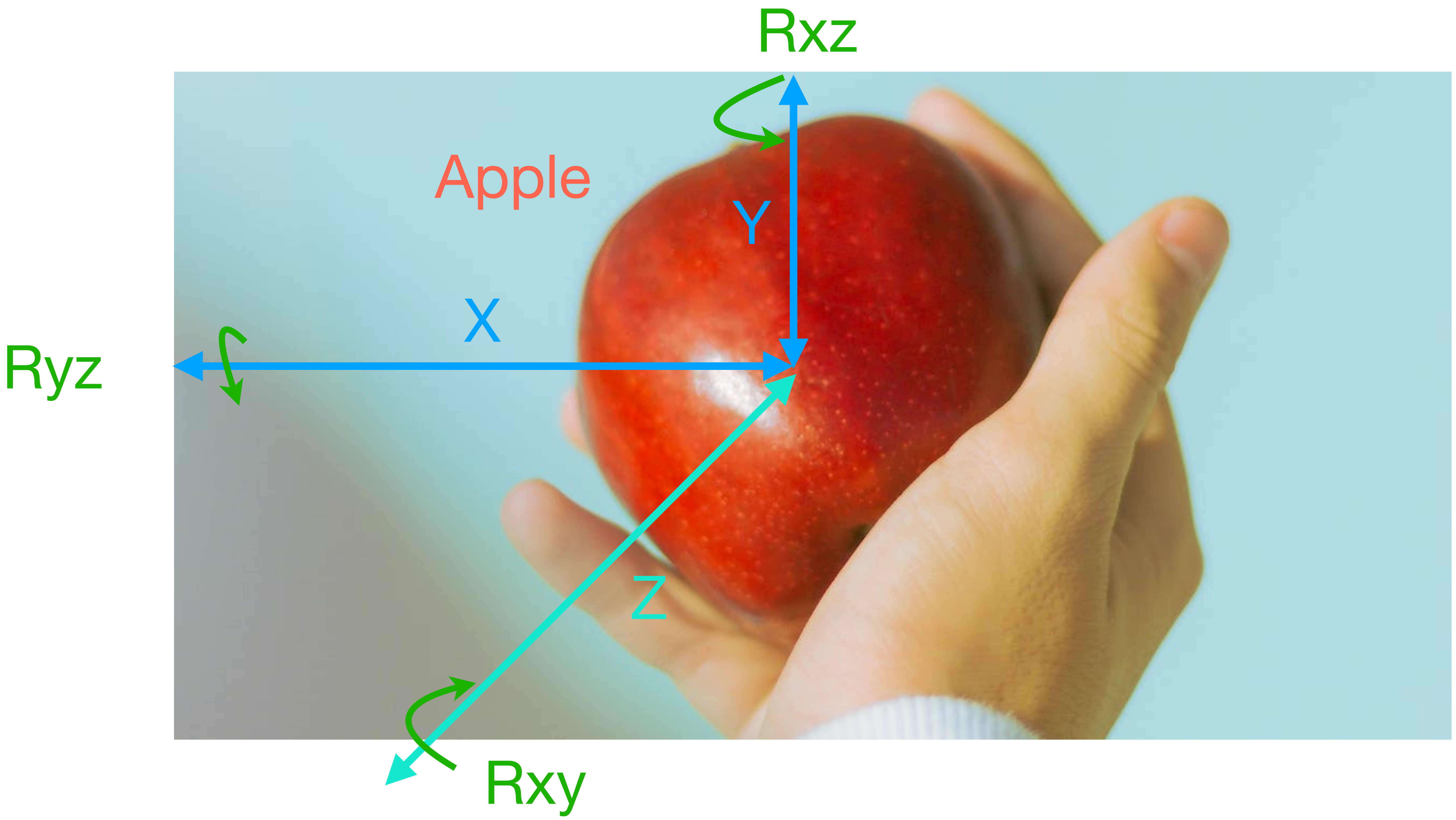
- When we see an object, we don't just see an abstract category.



Location Distance Rotation

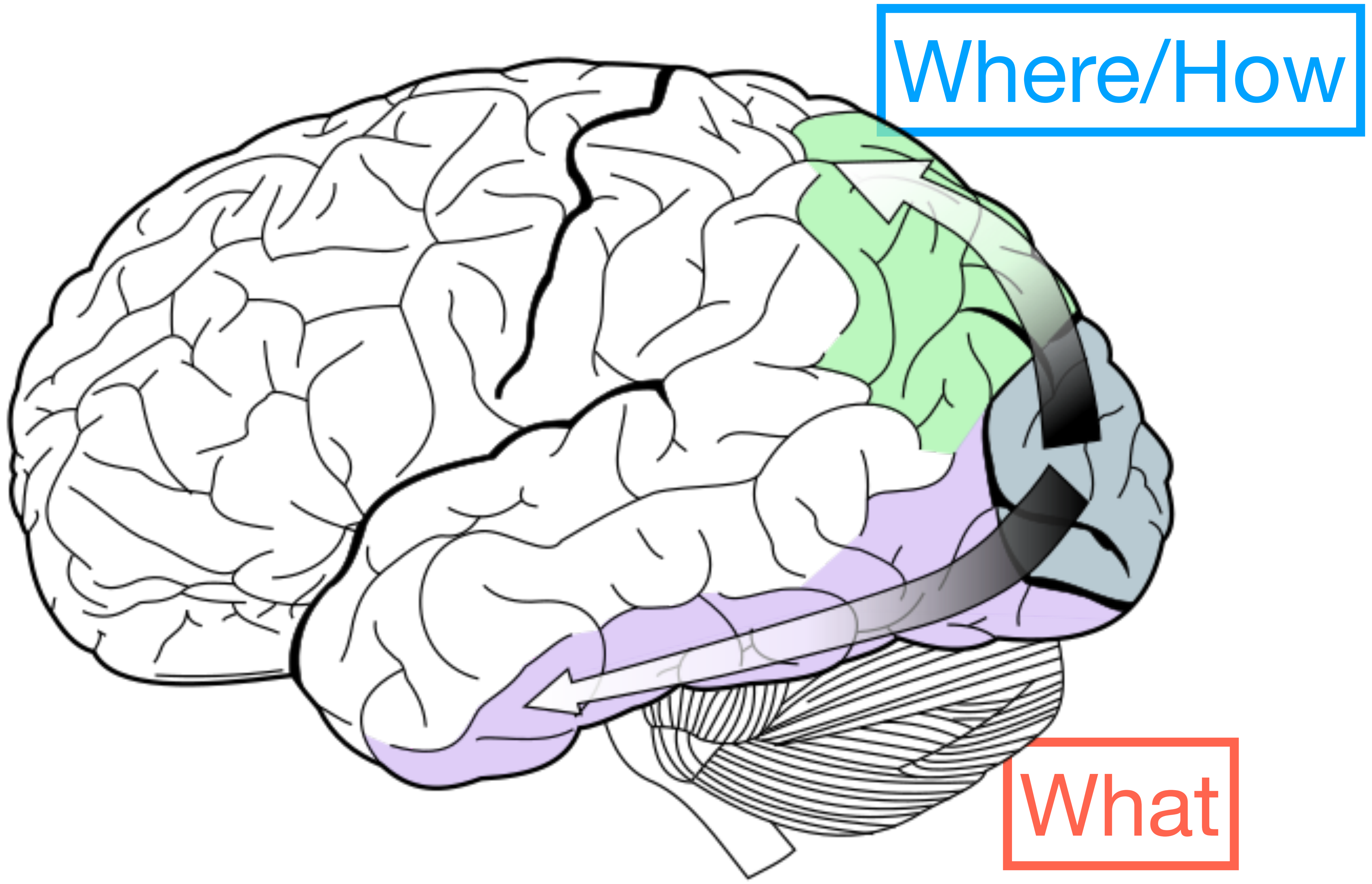
Vision is much more than object recognition

- When we see an object, we don't just see an abstract category.



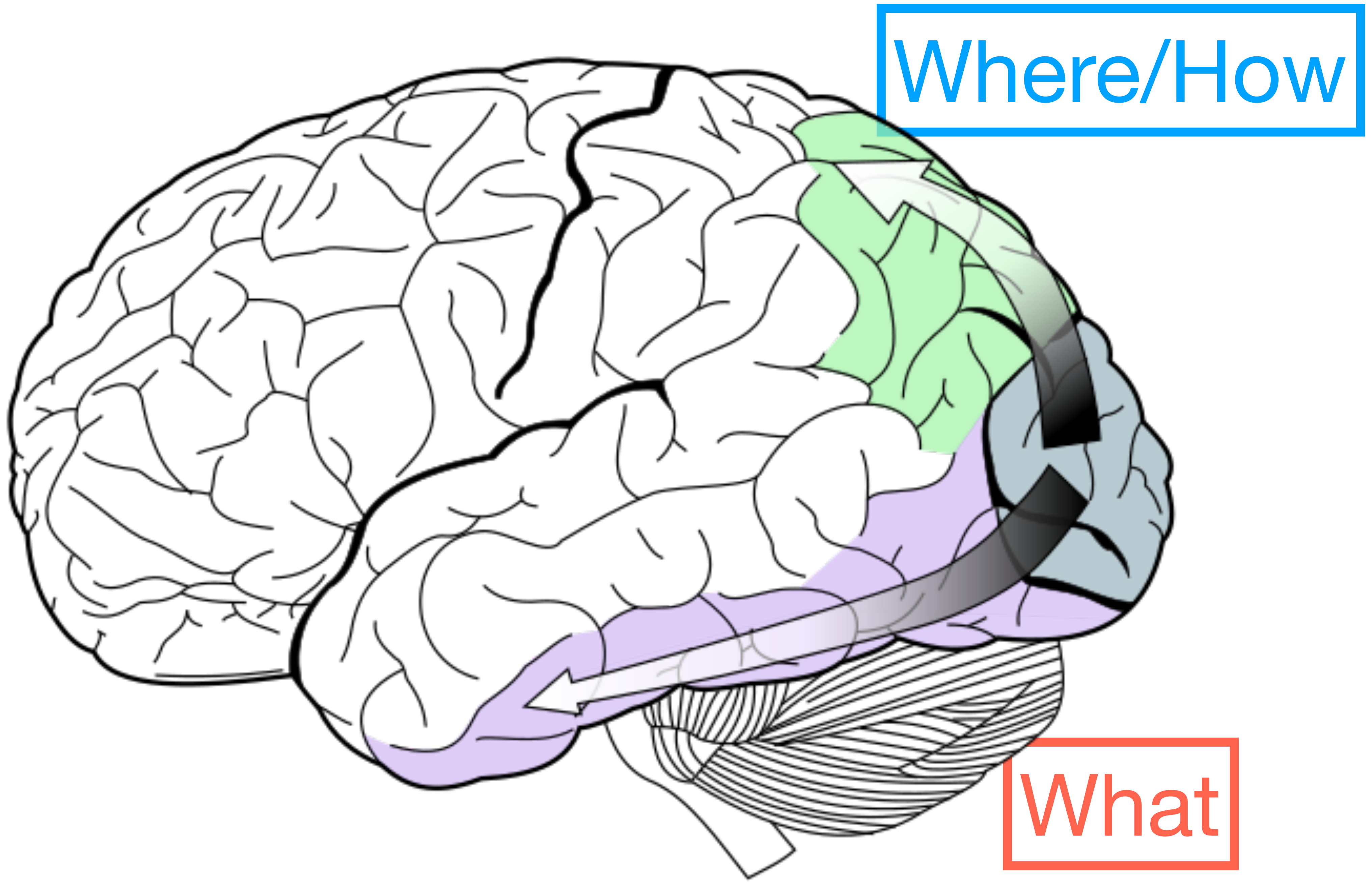
Location Distance Rotation Color, weight, etc.

Ventral stream is thought to perform the “what” function in vision



Brain schematic from wikipedia: two-streams hypothesis

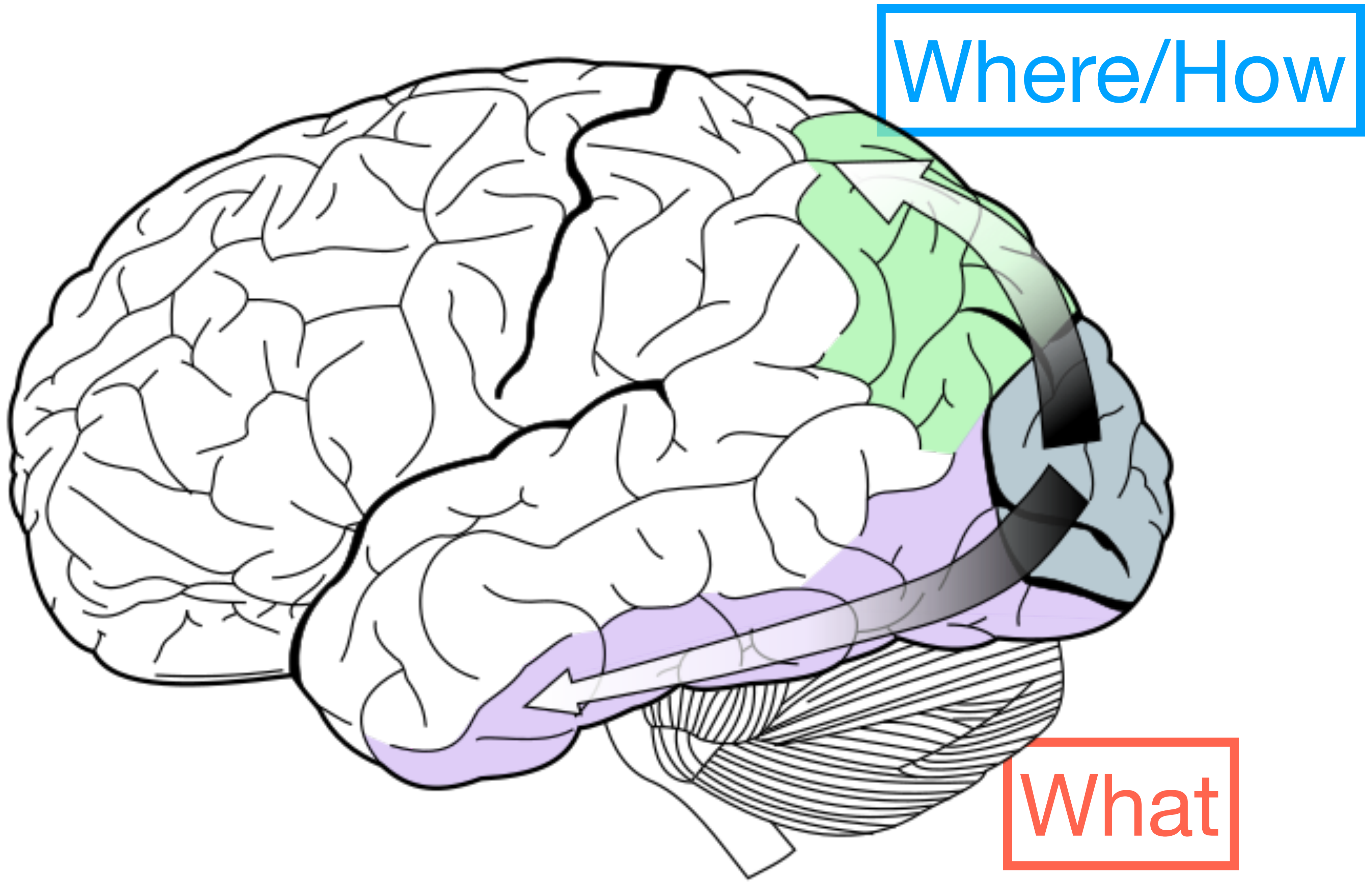
Ventral stream is thought to perform the "what" function in vision



- The ventral stream is thought to perform the "**What**" function in vision, while the dorsal stream performs spatial processing. Mishkin, Ungerleider (1982)

Brain schematic from wikipedia: two-streams hypothesis

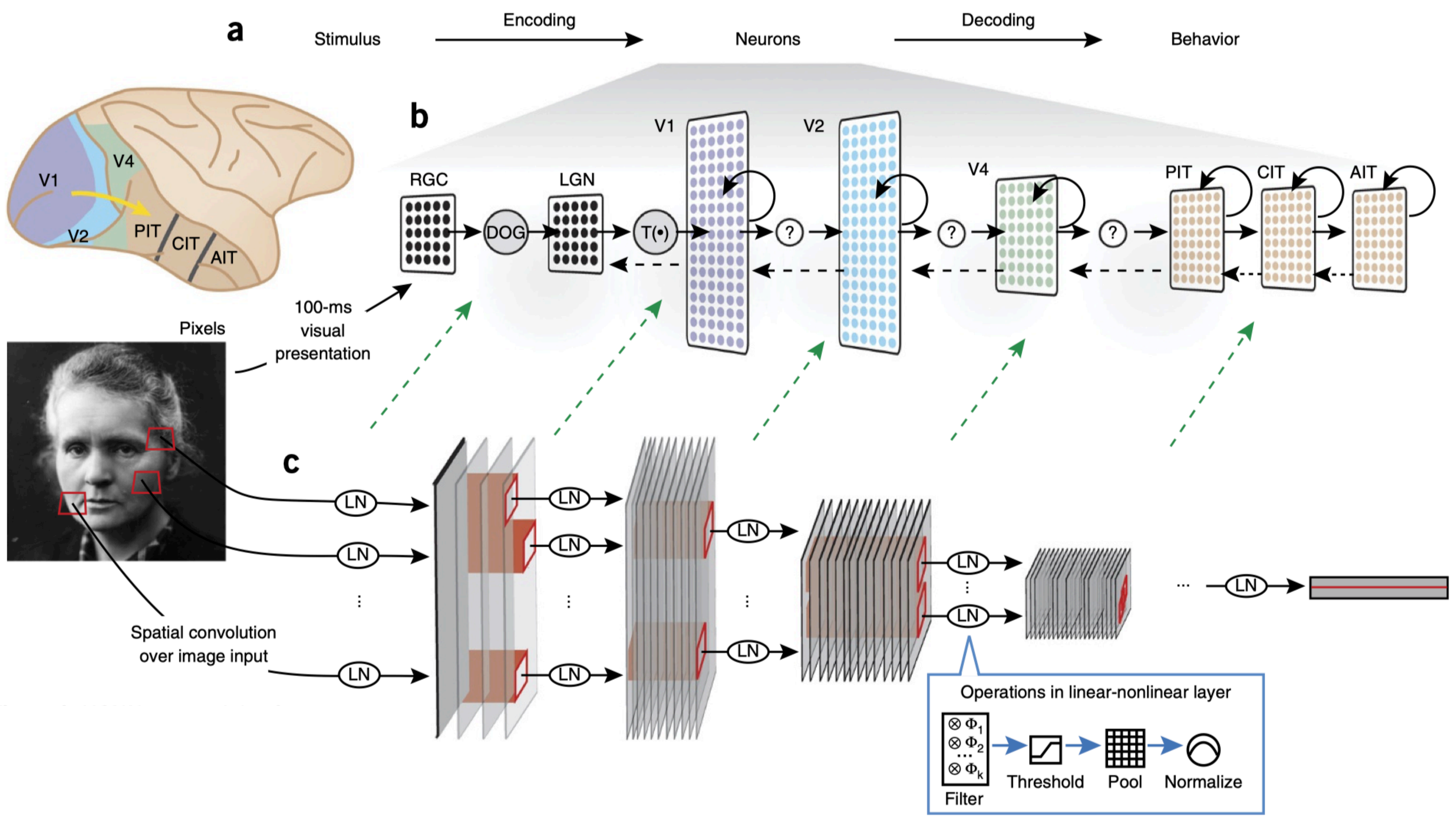
Ventral stream is thought to perform the “what” function in vision



- The ventral stream is thought to perform the “**What**” function in vision, while the dorsal stream performs spatial processing. Mishkin, Ungerleider (1982)
- Though, what exactly does the “what” function mean is debated. Goodale MA, Milner AD (1992)

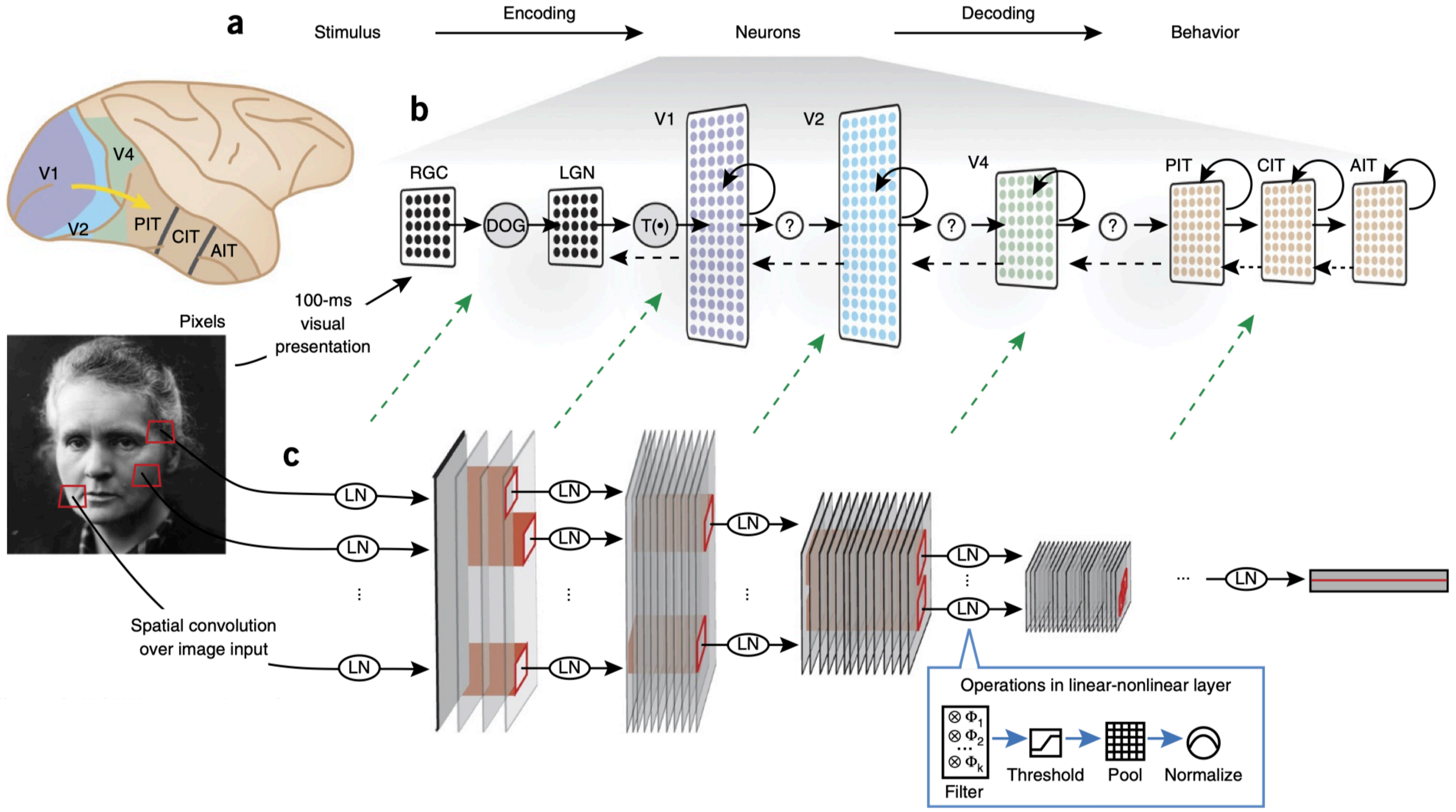
Brain schematic from wikipedia: two-streams hypothesis

Previous computational models are consistent with the “what” function



Yamins et al. 2016

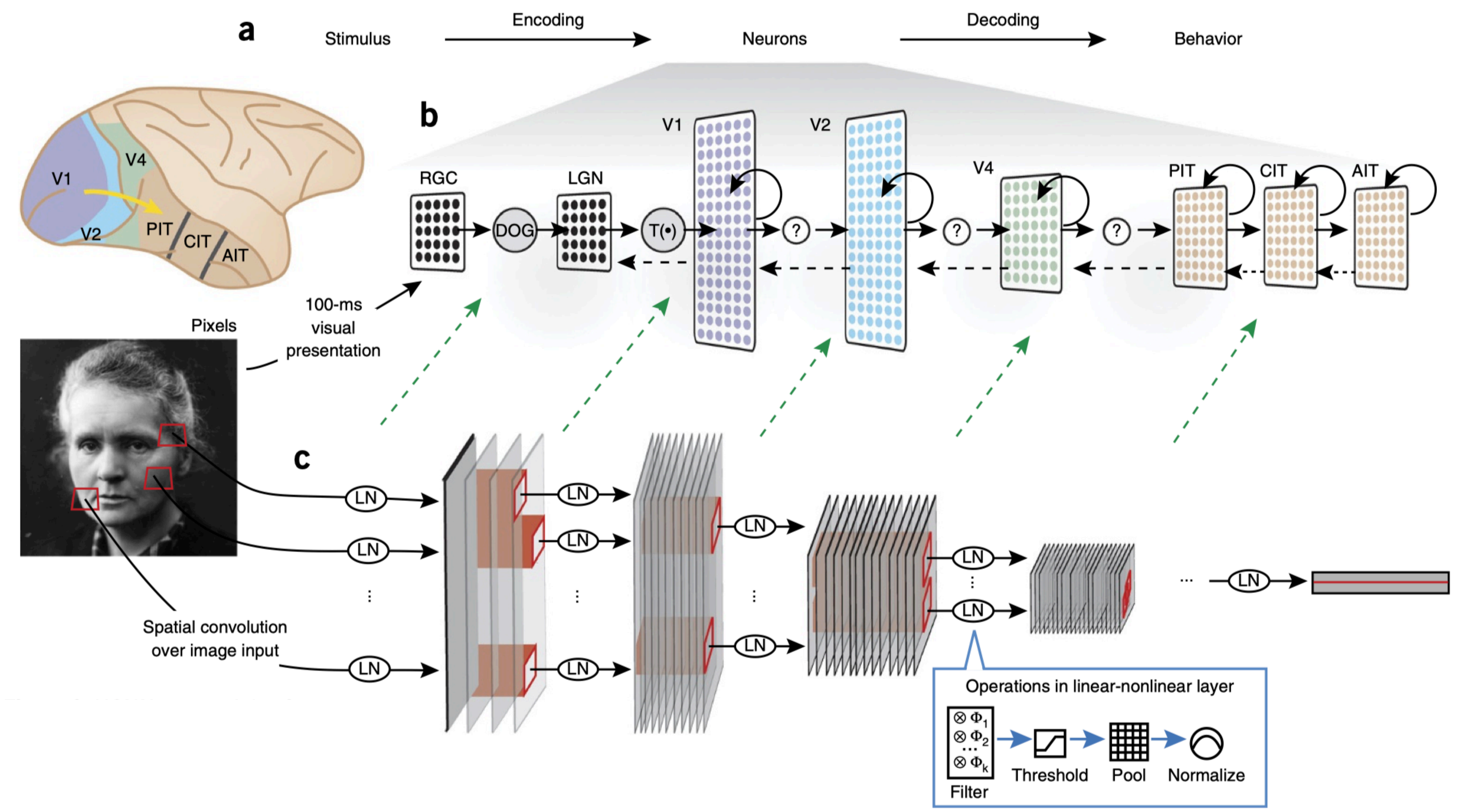
Previous computational models are consistent with the “what” function



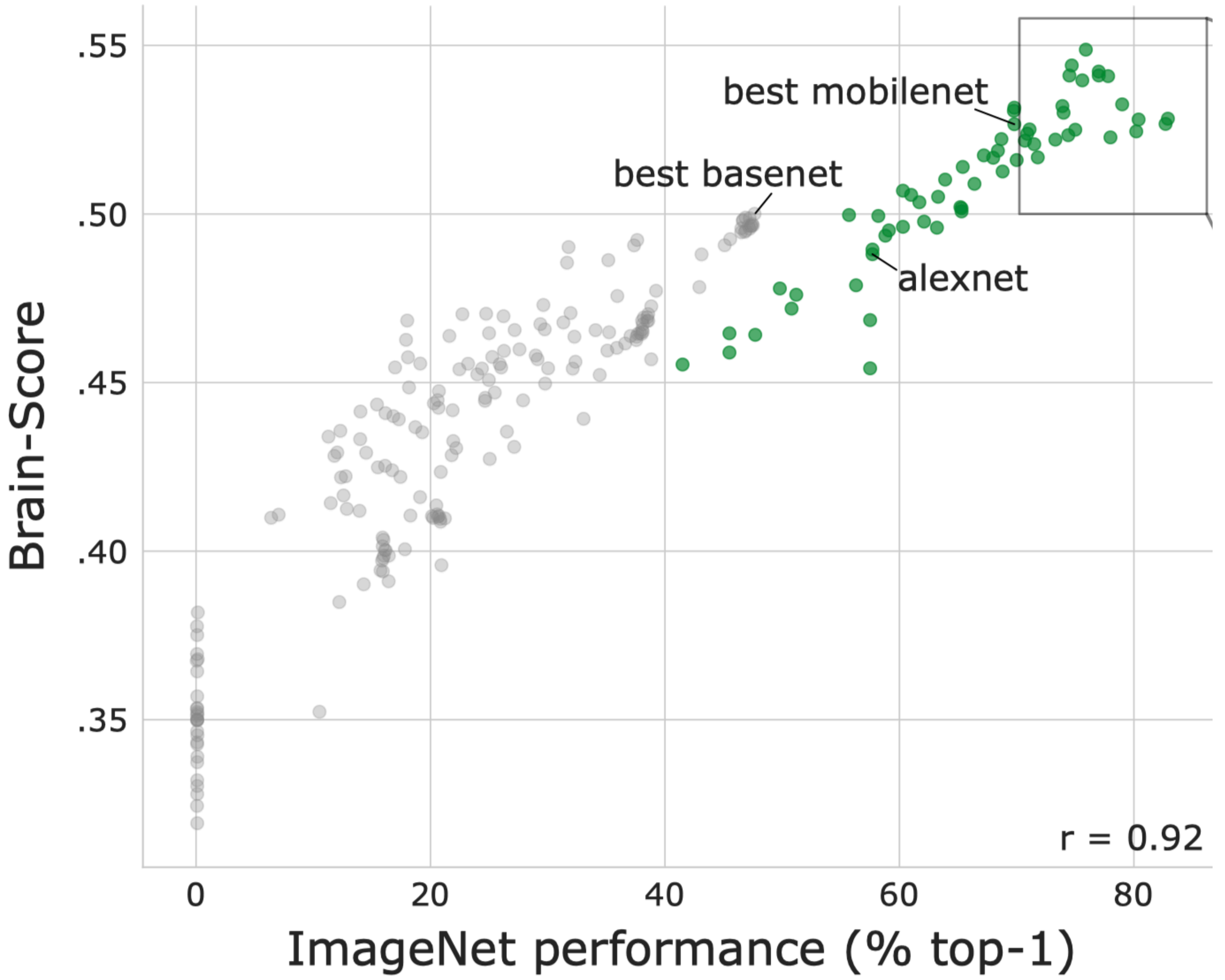
Yamins et al. 2016

- Most leading ventral stream models are derived by optimizing networks for **object categorization.**

Previous computational models are consistent with the "what" function



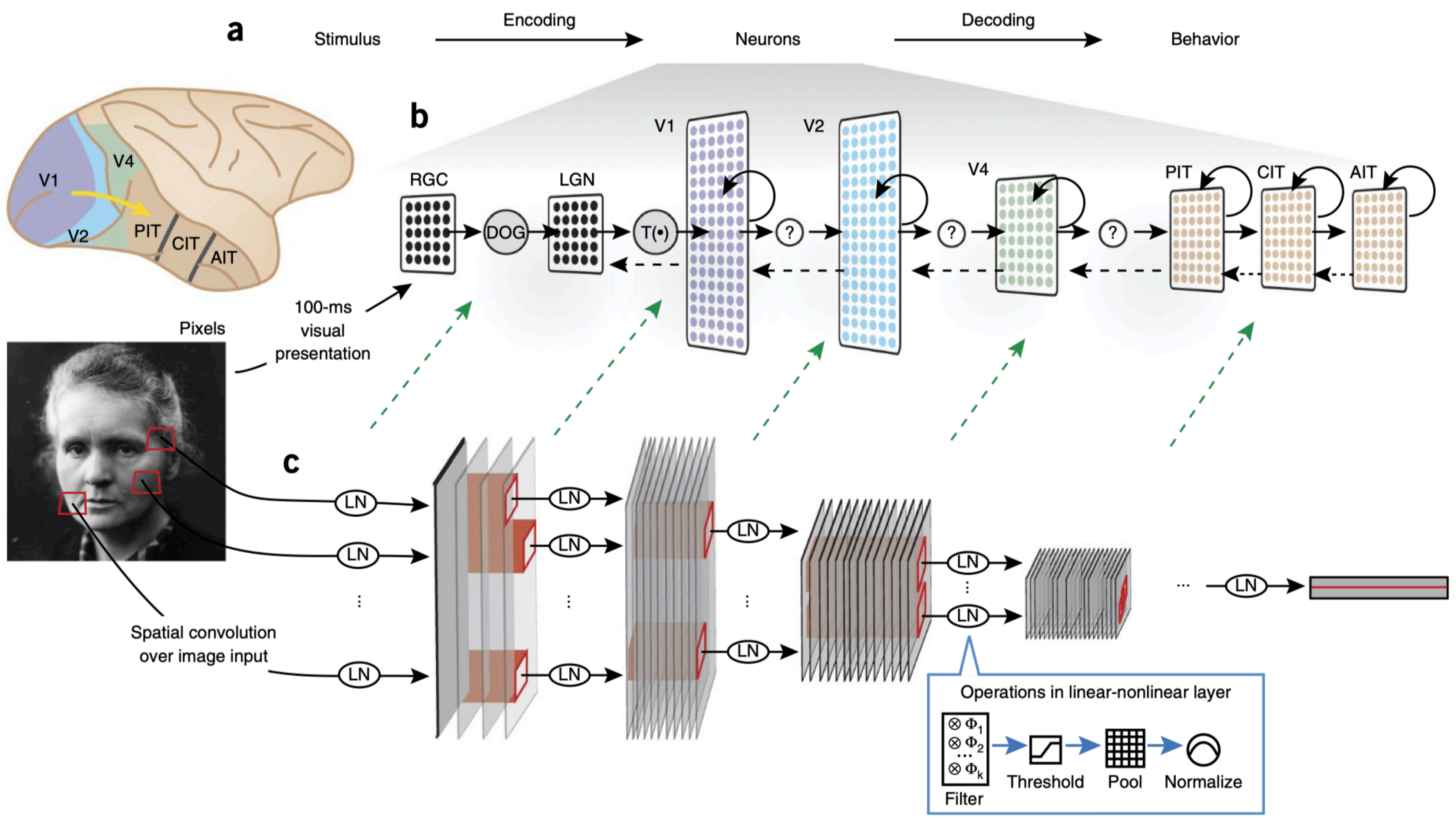
Yamins et al. 2016



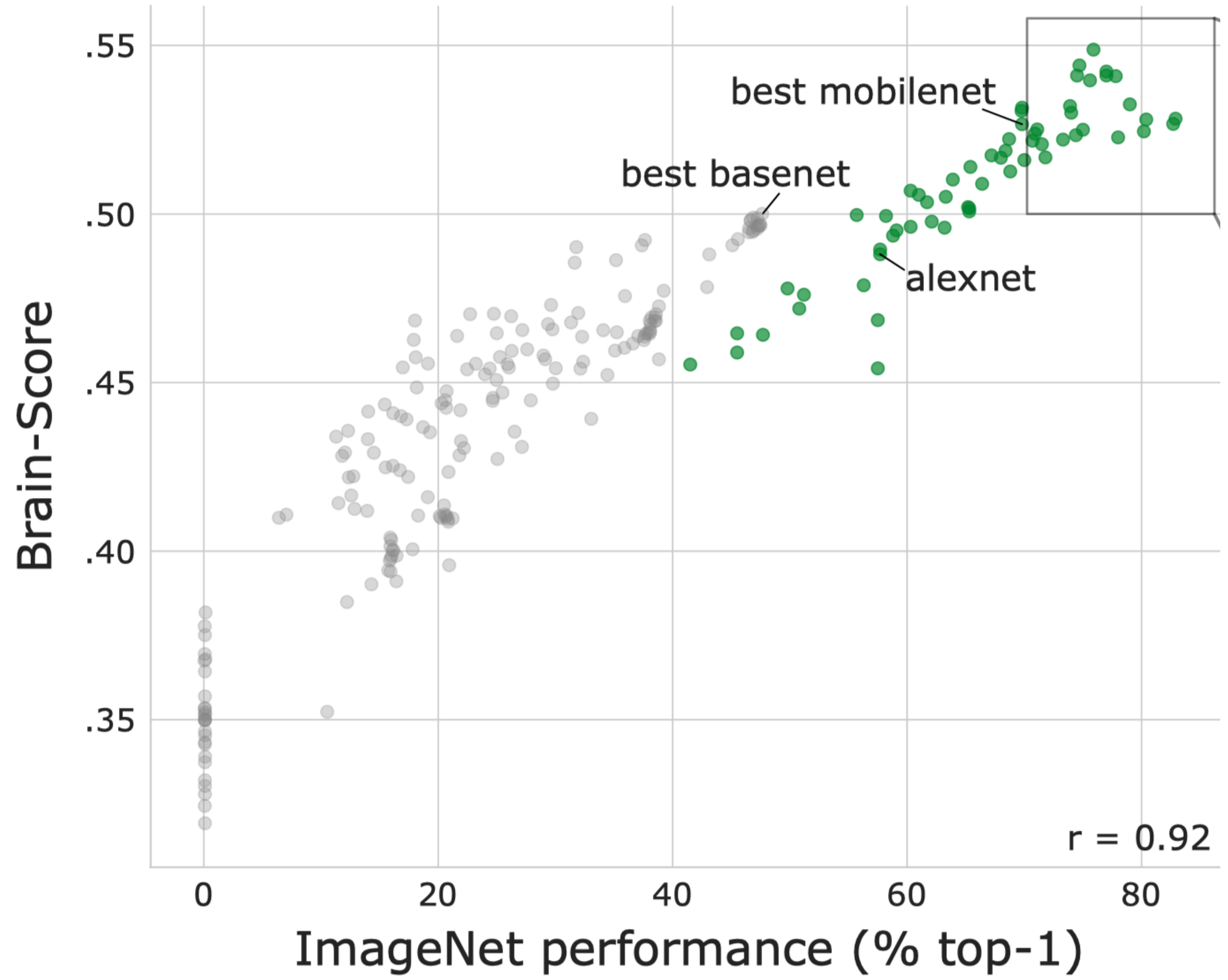
Schrimpf et al., 2020

- Most leading ventral stream models are derived by optimizing networks for **object categorization.**
- Categorization performance strongly correlates with ventral stream alignment.

Previous computational models are consistent with the “what” function



Yamins et al. 2016

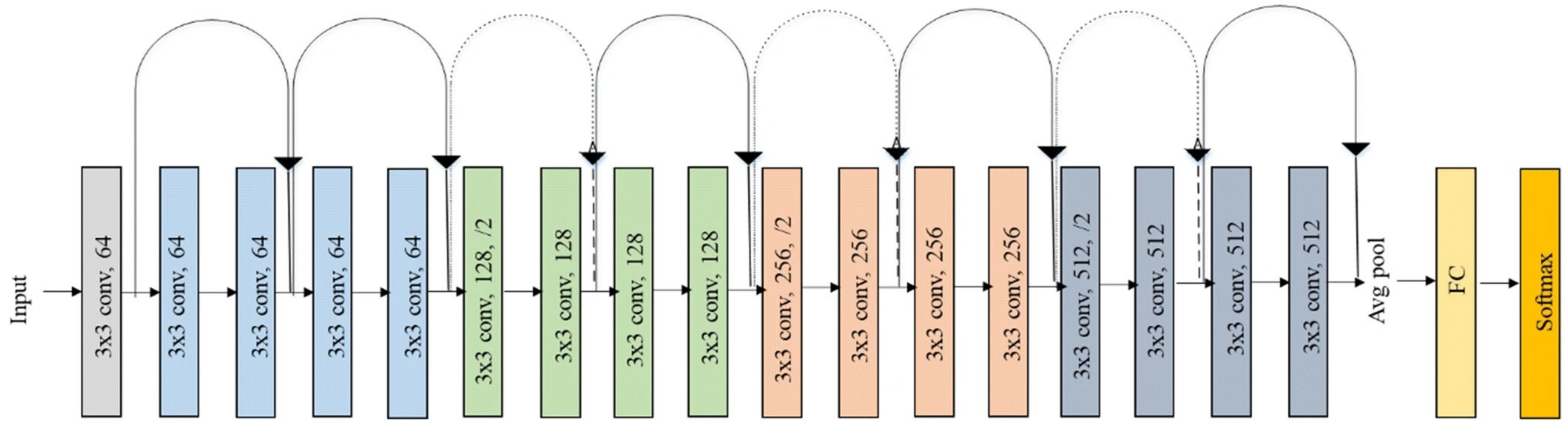


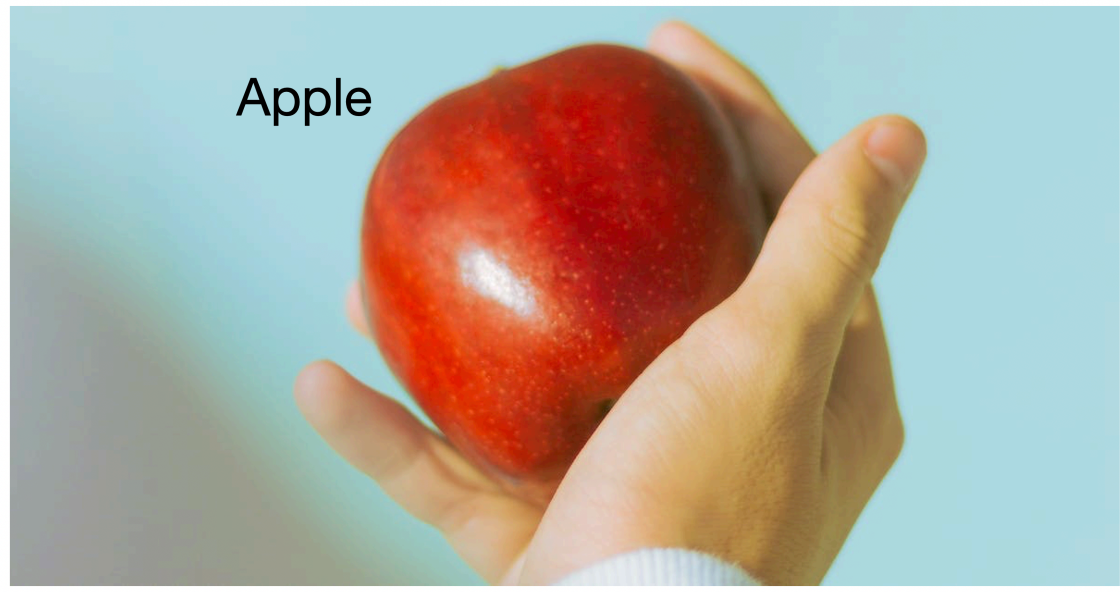
Schrimpf et al., 2020

- Most leading ventral stream models are derived by optimizing networks for **object categorization.**
- Categorization performance strongly correlates with ventral stream alignment.
- This seems to imply that evolution/development derived the ventral stream under the objective of object categorization.

Might the ventral stream be optimized for spatial tasks?

CNN trained to estimate image latents using supervised learning

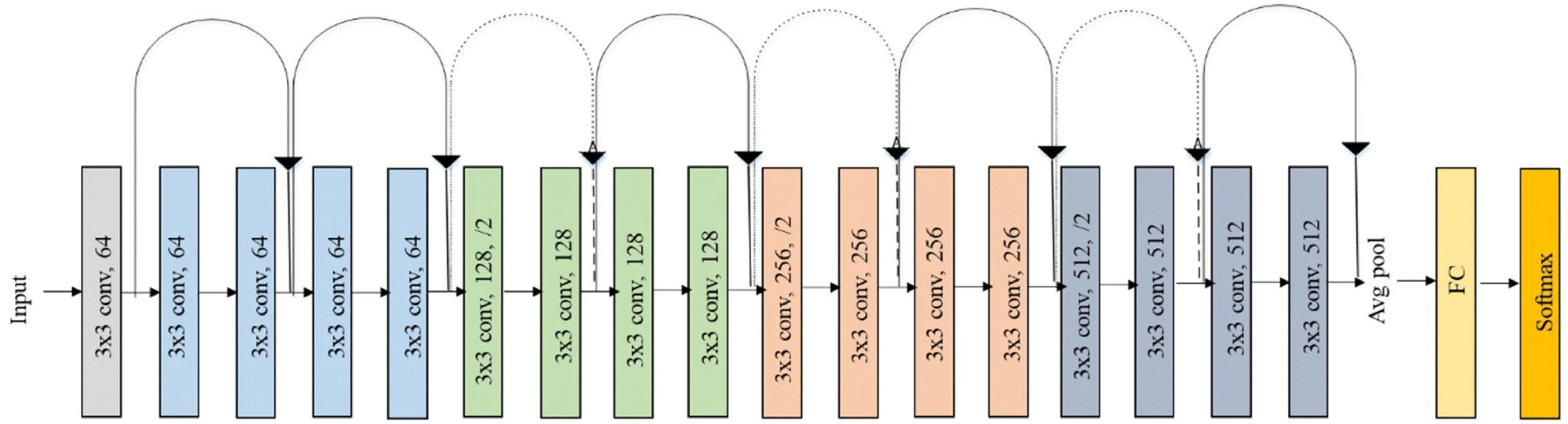


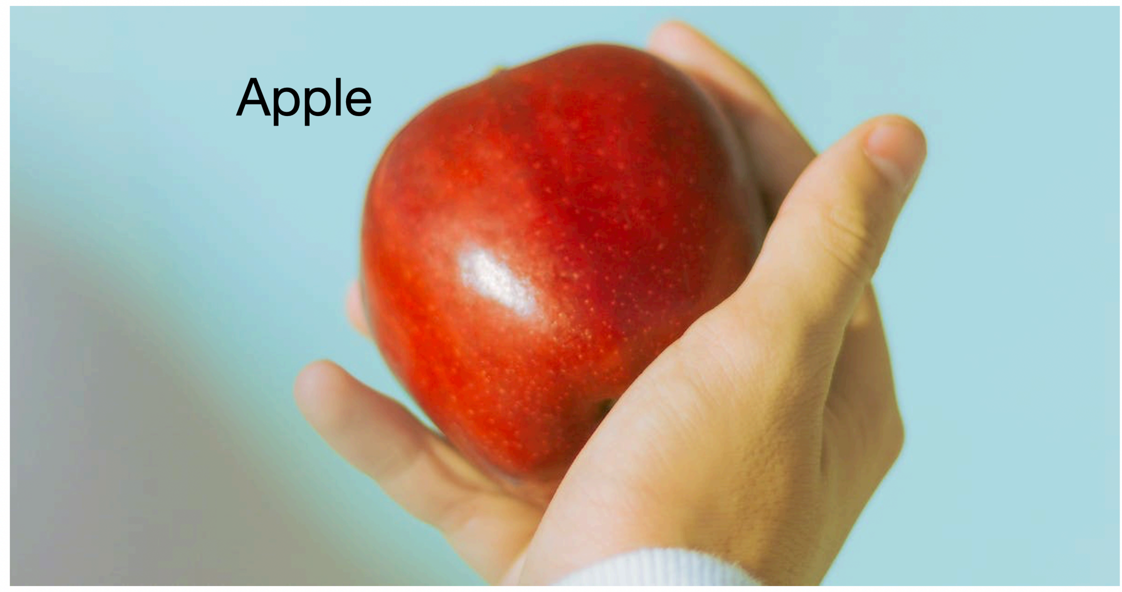


Category

CNN schematic from Ramzan, et al., 2020

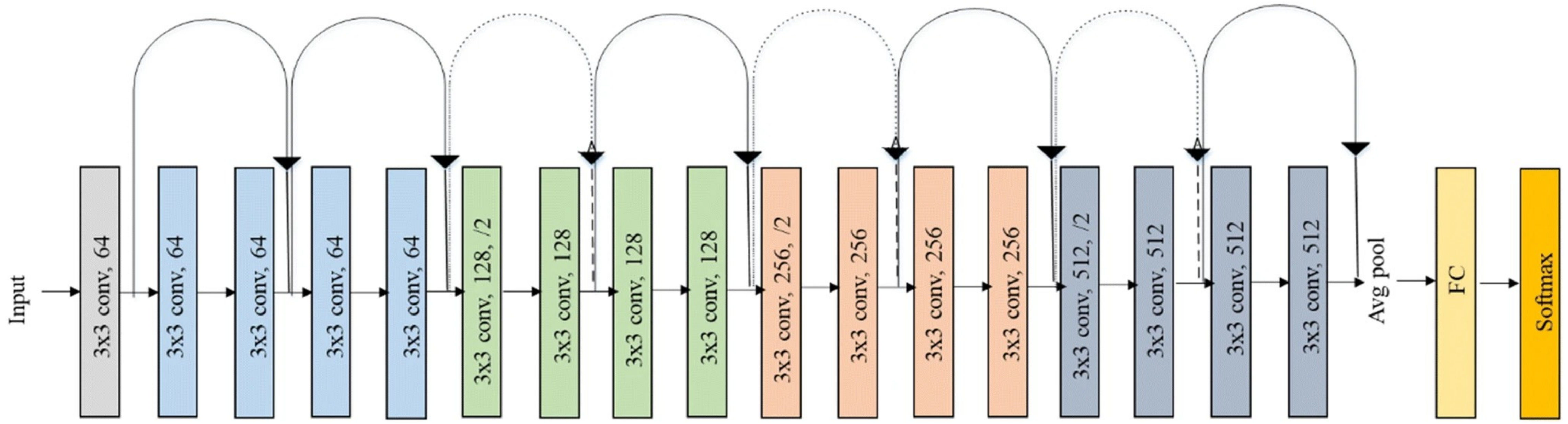
CNN trained to estimate image latents using supervised learning

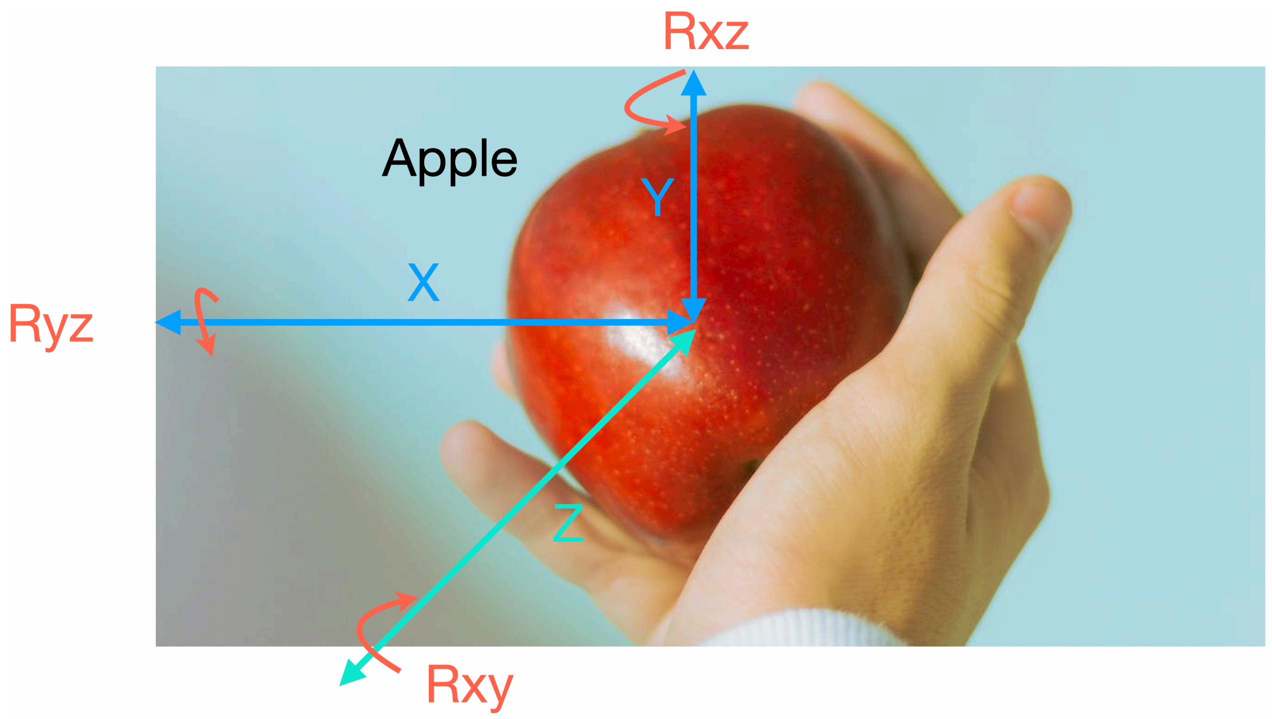




Category

Vs





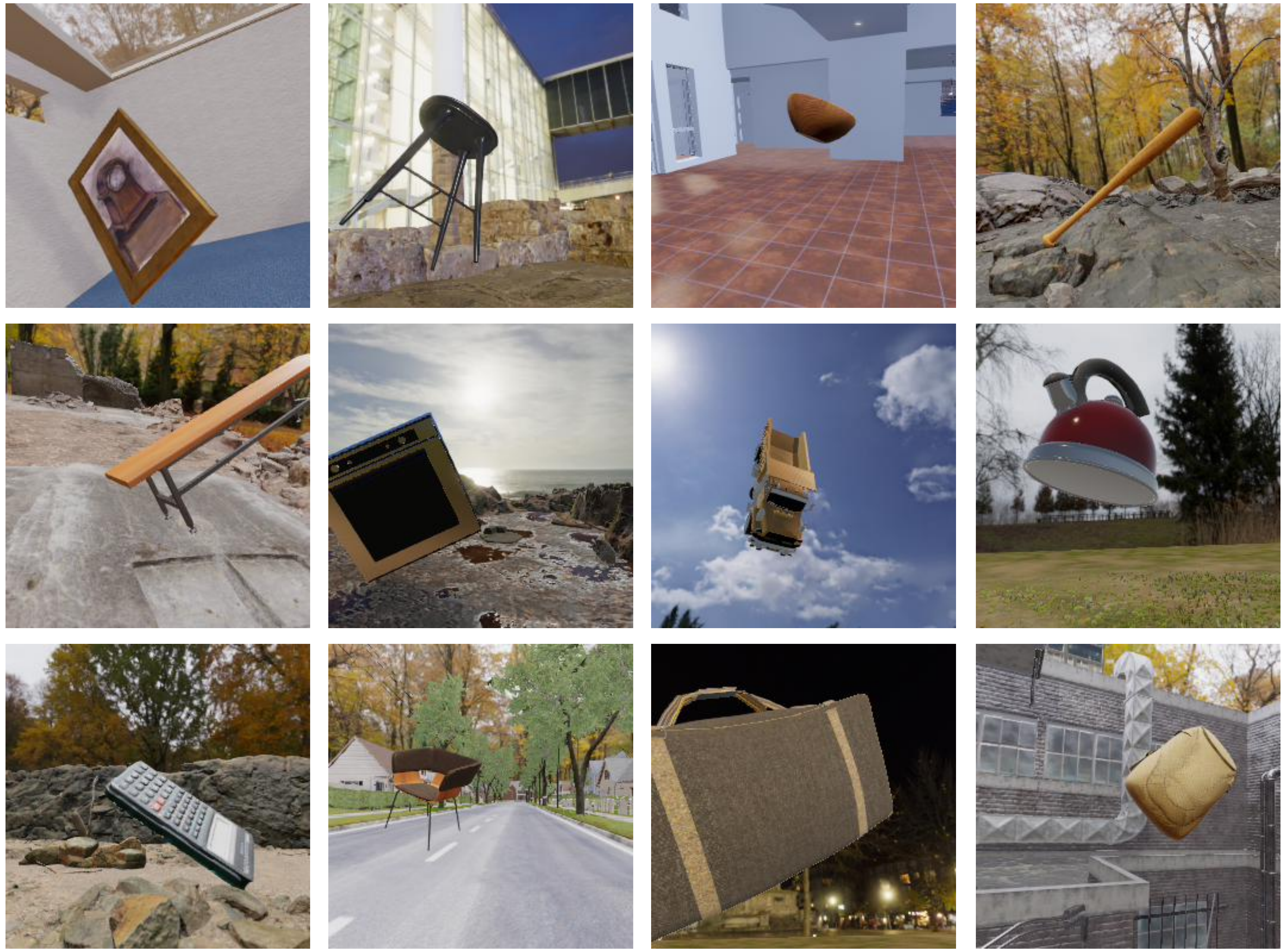
Location, Distance, Rotation

...

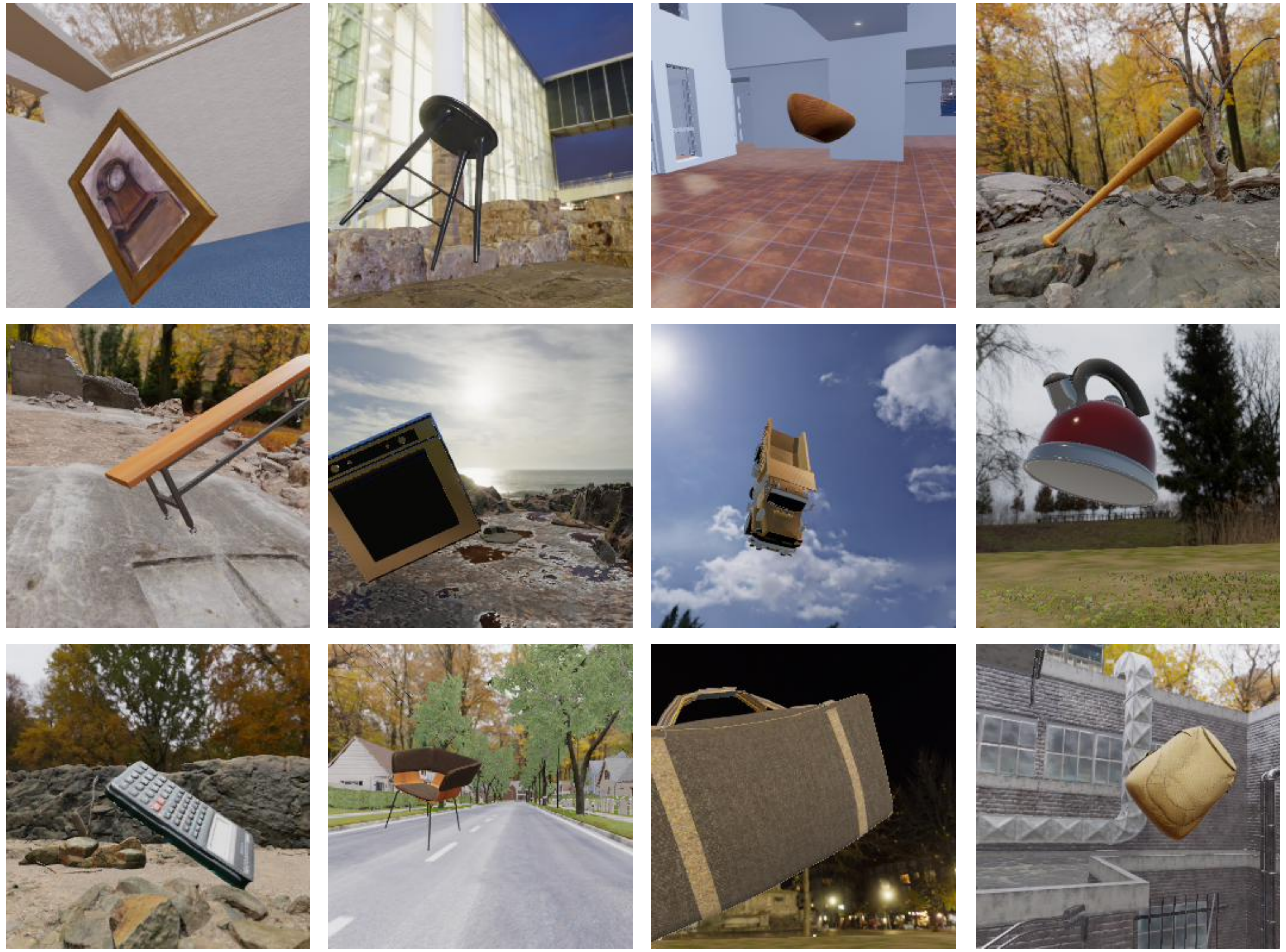
CNN schematic from Ramzan, et al., 2020

Synthetic image datasets that contain rich ground-truth latent labels

TDW ThreeDWorld

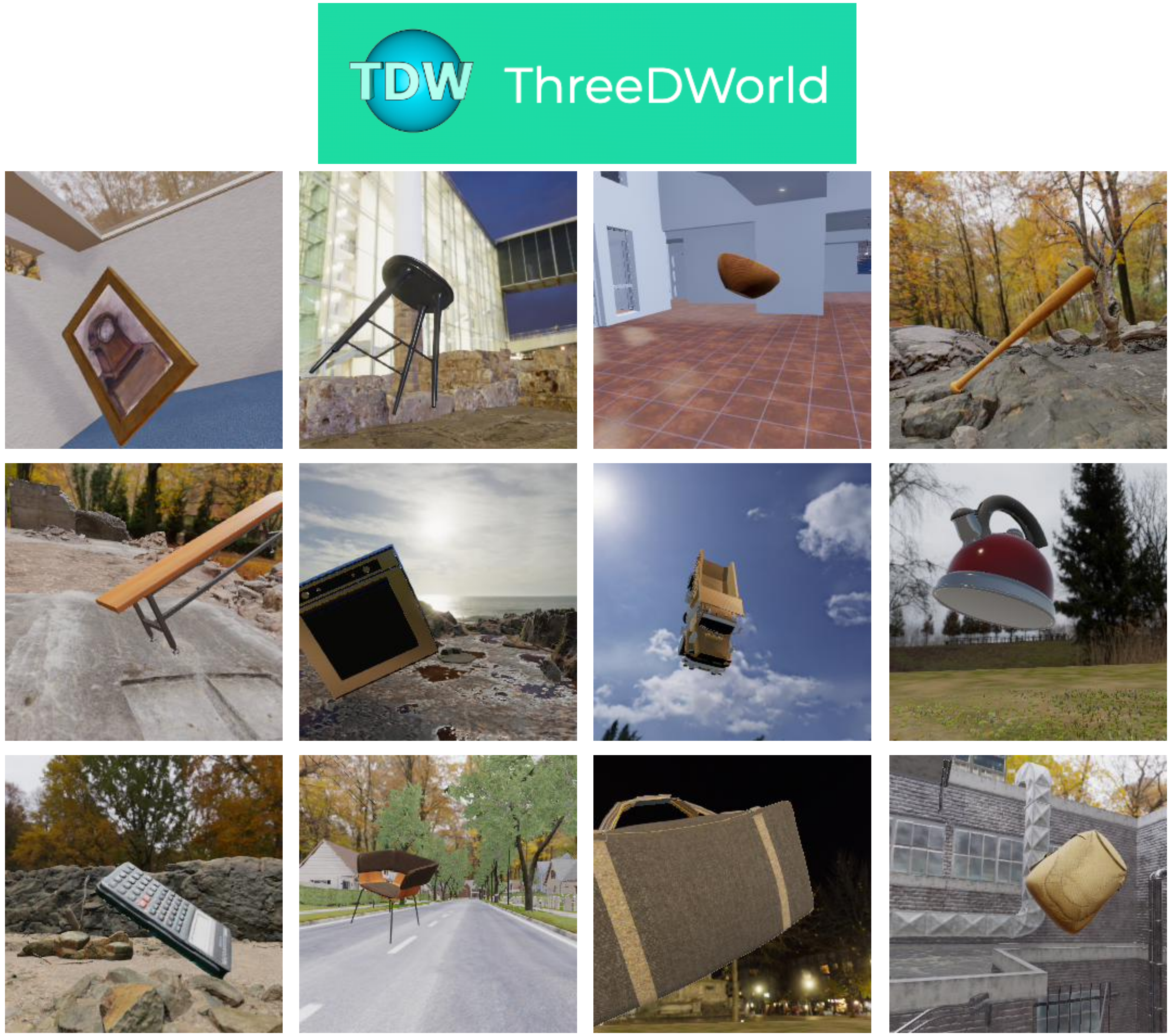


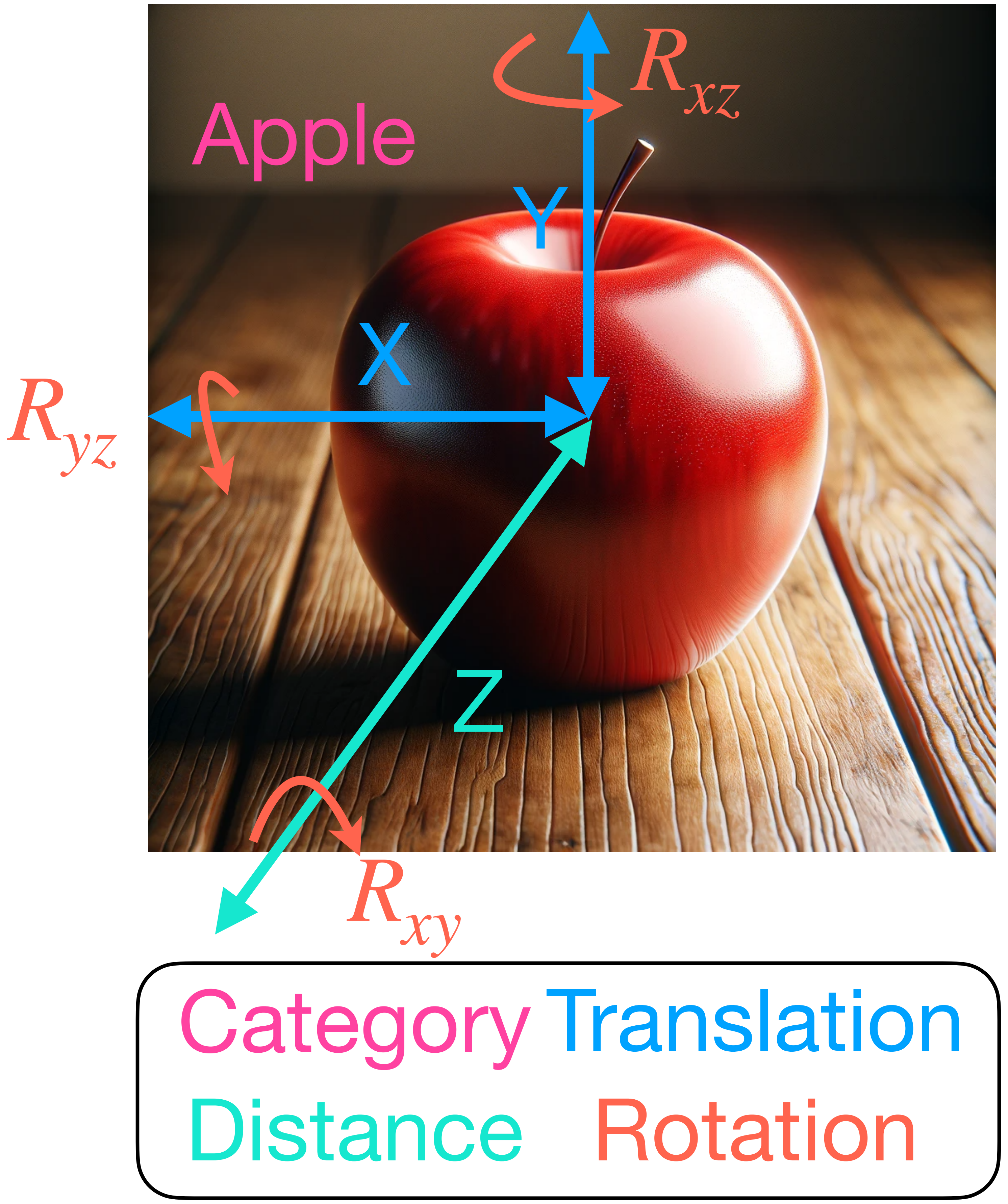
Synthetic image datasets that contain rich ground-truth latent labels



- We used TDW, a 3D graphic engine, to generate large image datasets (up to100M) that contain rich ground-truth latent labels.

Synthetic image datasets that contain rich ground-truth latent labels

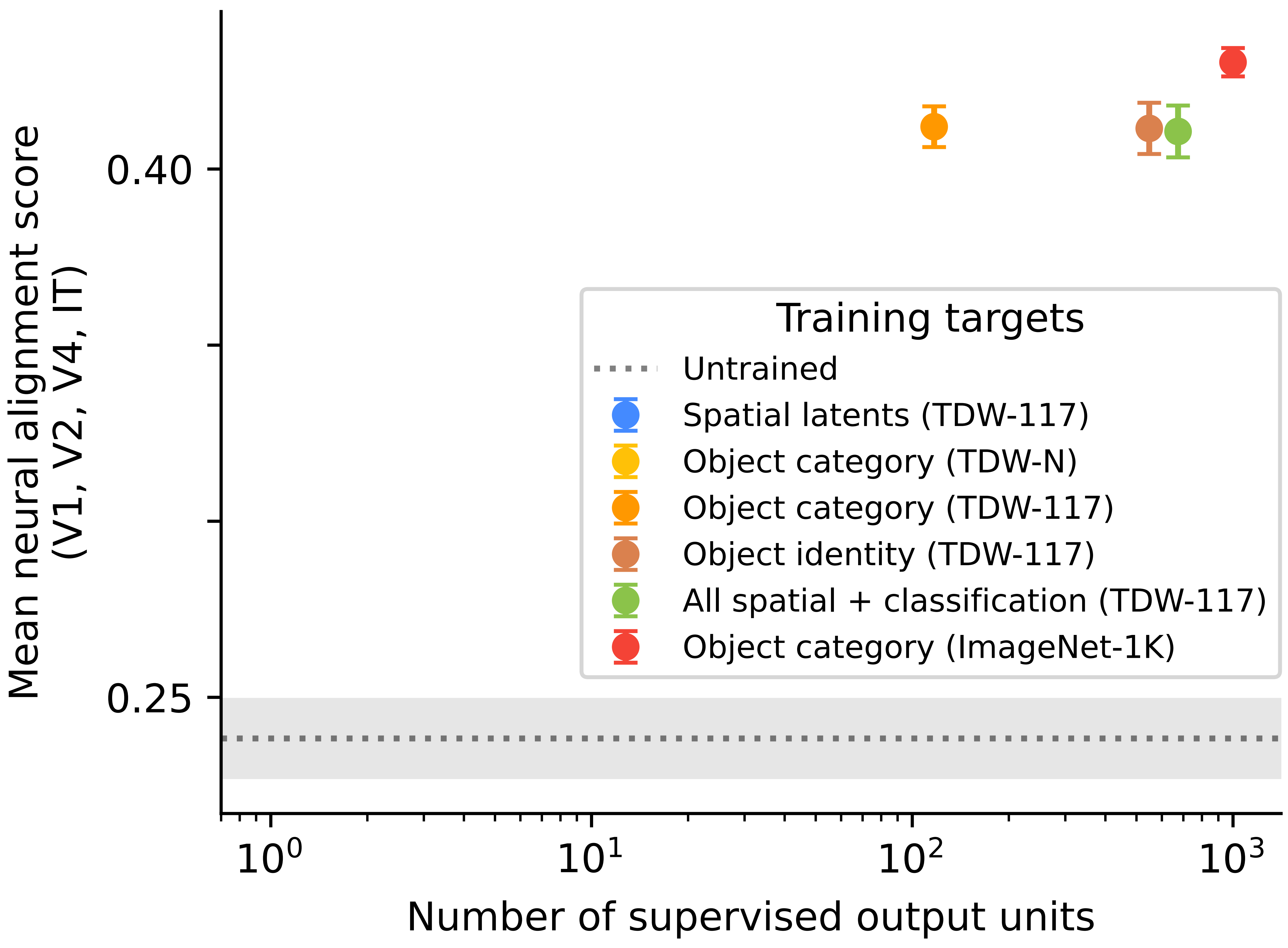




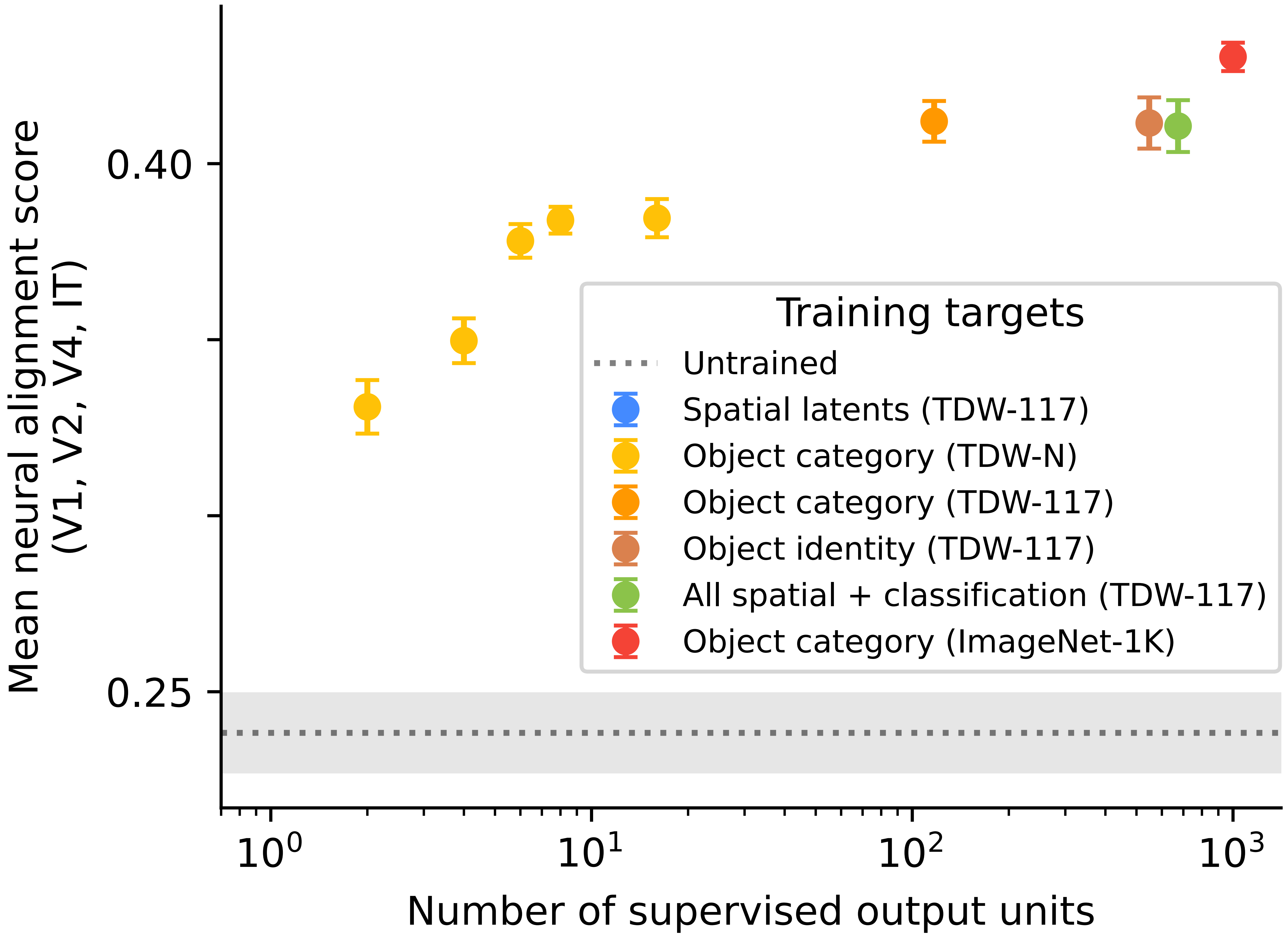
- We used TDW, a 3D graphic engine, to generate large image datasets (up to100M) that contain rich ground-truth latent labels.

Learning a few spatial latents produces ventral-stream-aligned CNN

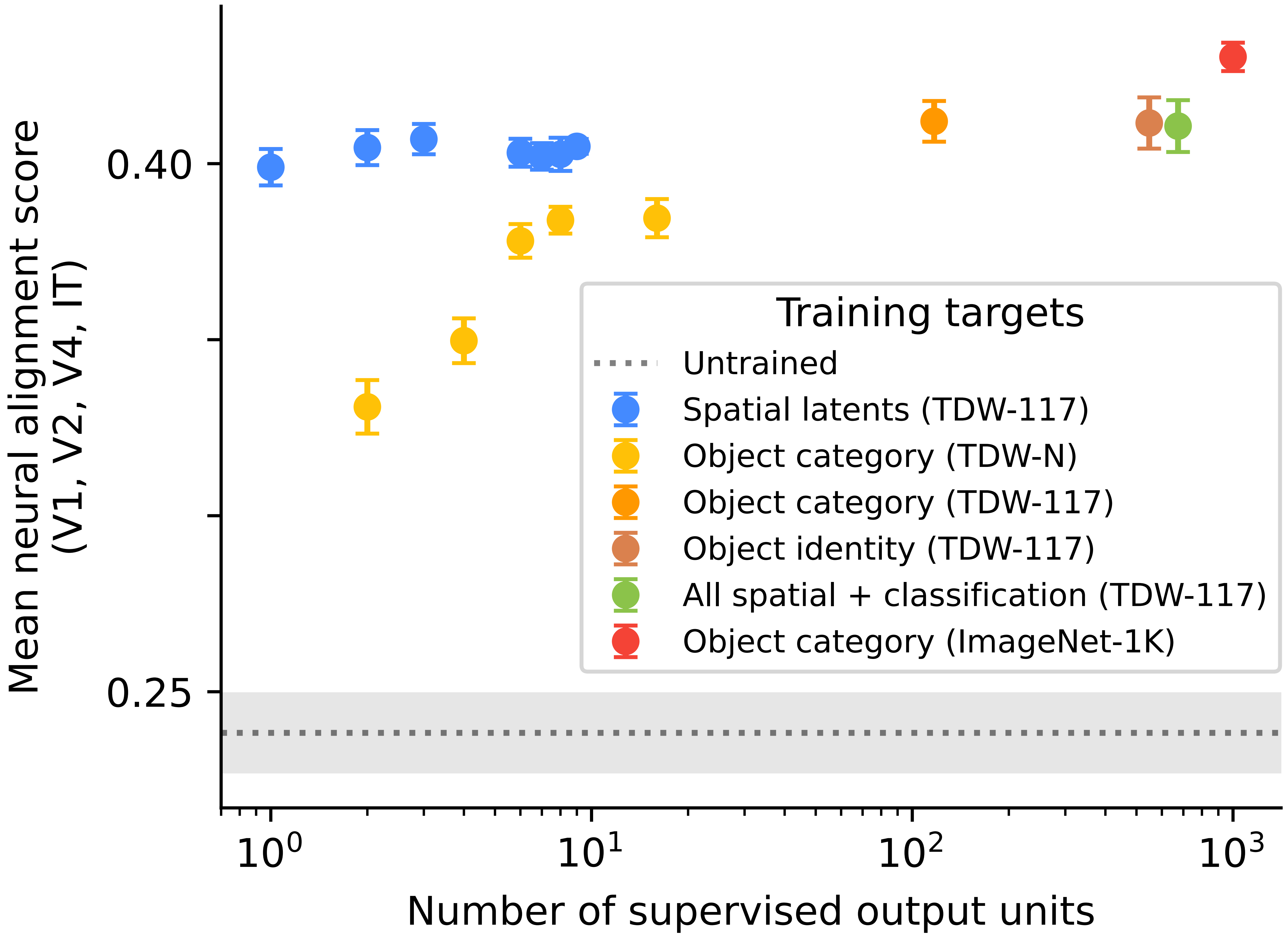
Learning a few spatial latents produces ventral-stream-aligned CNN



Learning a few spatial latents produces ventral-stream-aligned CNN



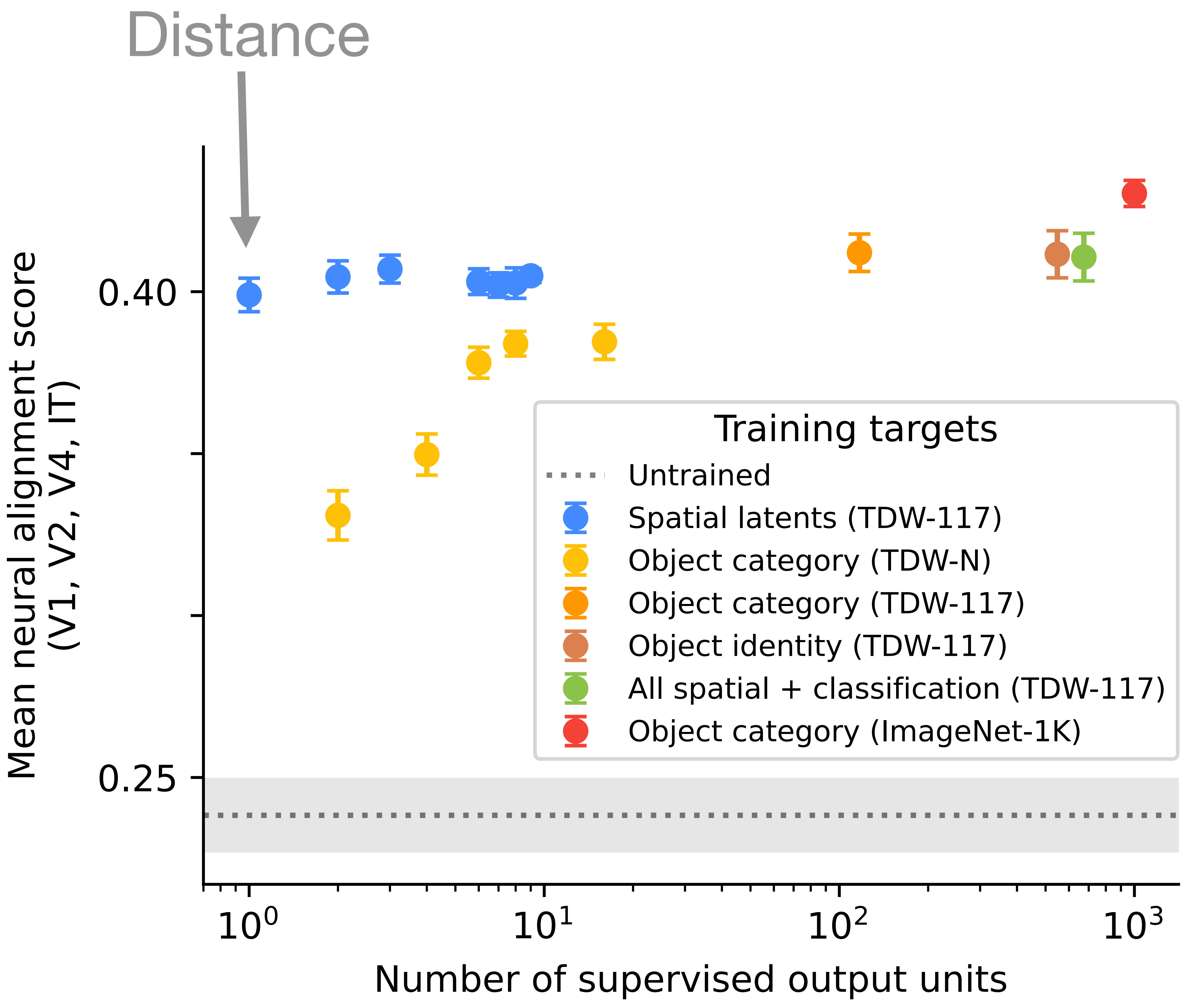
Learning a few spatial latents produces ventral-stream-aligned CNN



	Training task	# output targets
Spatial latent regression (TDW-117)	Distance regression	1
	Translation regression	2
	Distance + Translation	3
	Rotation regression	6
	Distance + Rotation	7
	Translation + Rotation	8
	Distance + Translation + Rotation	9
Classification (TDW-117)	Object category classification	117
	Object identity classification	548
	All spatial latents + classification	674
Reference	Untrained	NA
	ImageNet-1K classification	1000

- CNNs trained on a few latents are comparable to those trained on hundreds of categories.

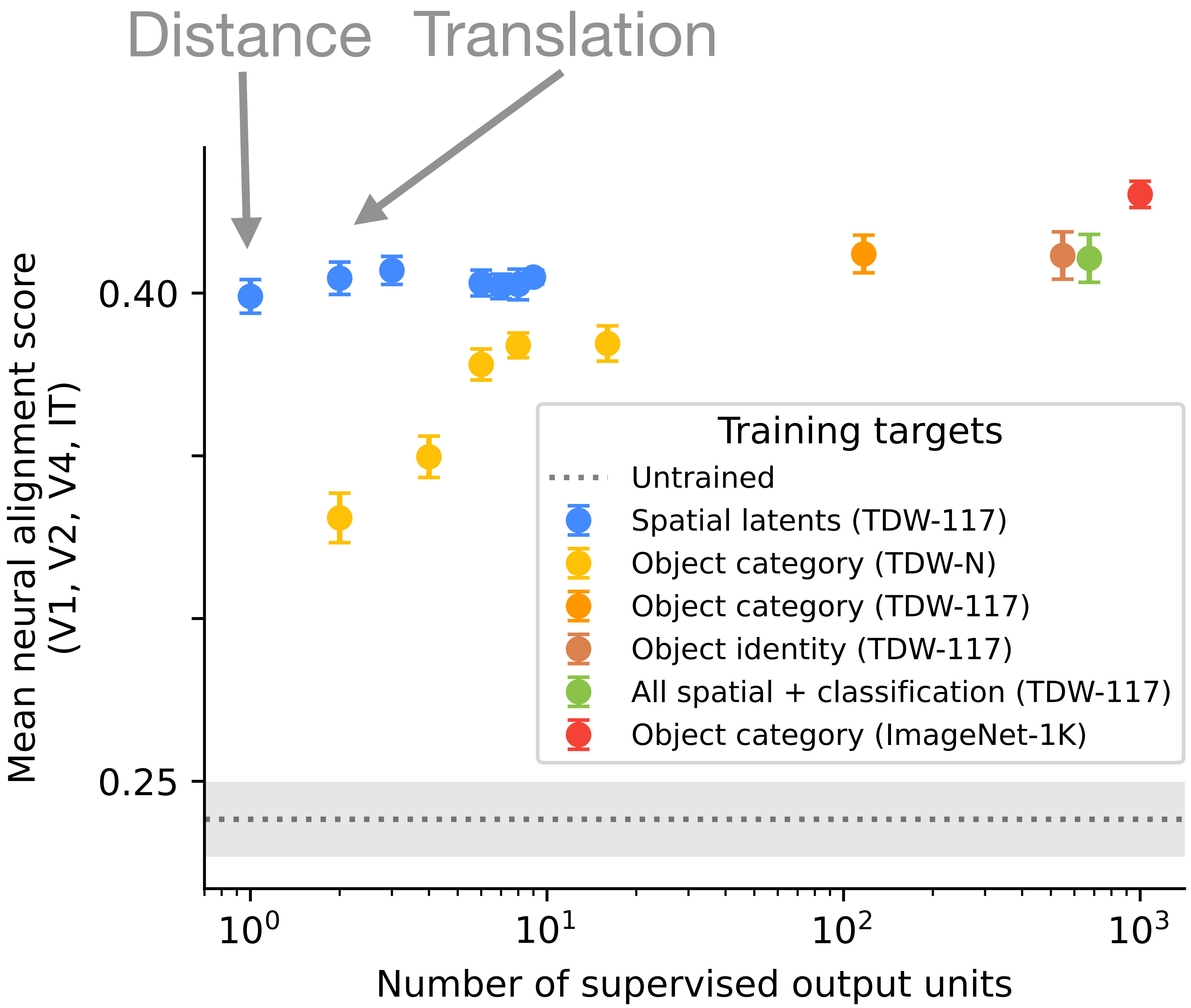
Learning a few spatial latents produces ventral-stream-aligned CNN



	Training task	**# output targets**
Spatial latent regression (TDW-117)	Distance regression	1
	Translation regression	2
	Distance + Translation	3
	Rotation regression	6
	Distance + Rotation	7
	Translation + Rotation	8
	Distance + Translation + Rotation	9
Classification (TDW-117)	Object category classification	117
	Object identity classification	548
	All spatial latents + classification	674
Reference	Untrained	NA
	ImageNet-1K classification	1000

- CNNs trained on a few latents are comparable to those trained on hundreds of categories.

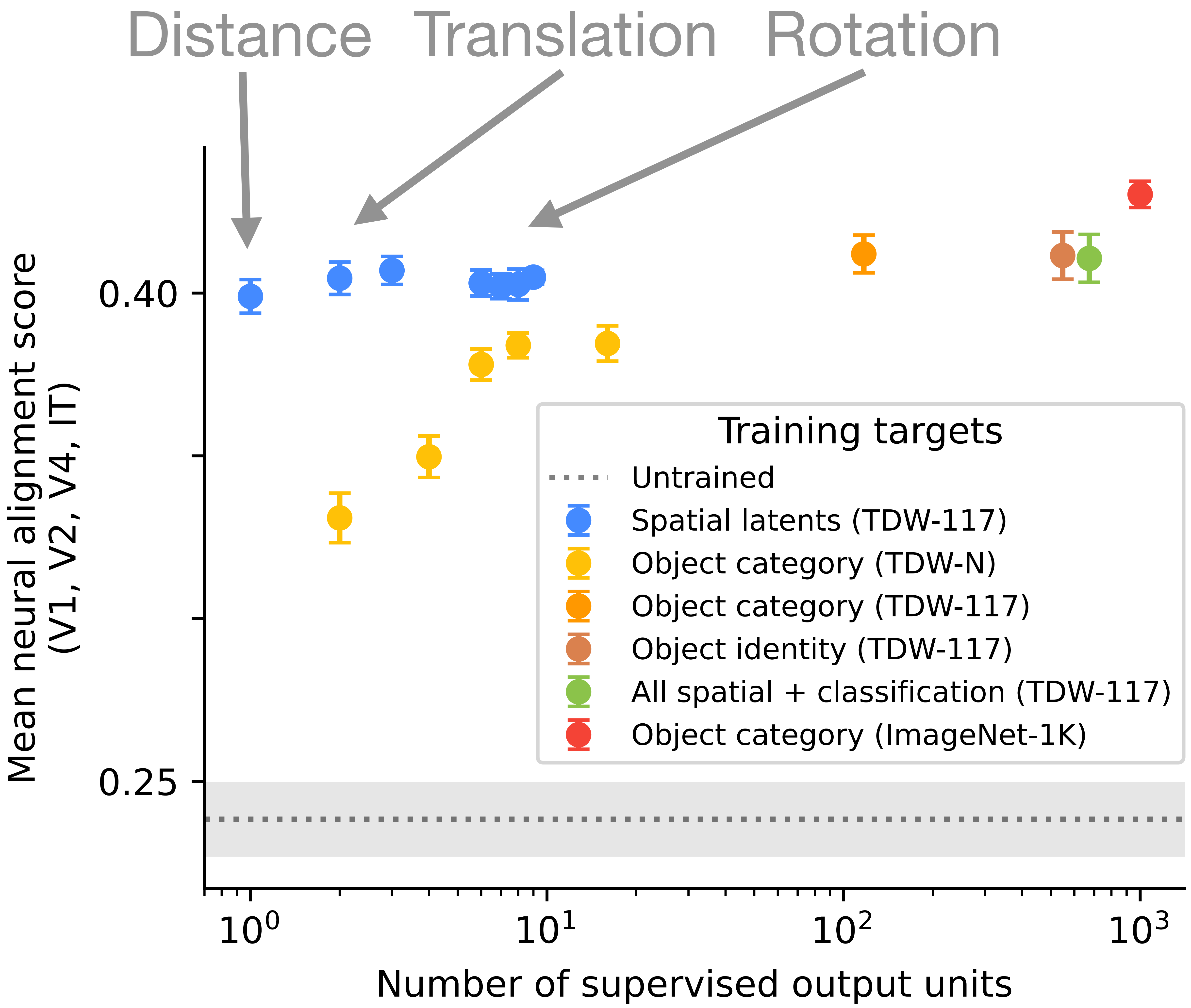
Learning a few spatial latents produces ventral-stream-aligned CNN



	Training task	# output targets
Spatial latent regression (TDW-117)	Distance regression	1
	Translation regression	2
	Distance + Translation	3
	Rotation regression	6
	Distance + Rotation	7
	Translation + Rotation	8
	Distance + Translation + Rotation	9
Classification (TDW-117)	Object category classification	117
	Object identity classification	548
	All spatial latents + classification	674
Reference	Untrained	NA
	ImageNet-1K classification	1000

- CNNs trained on a few latents are comparable to those trained on hundreds of categories.

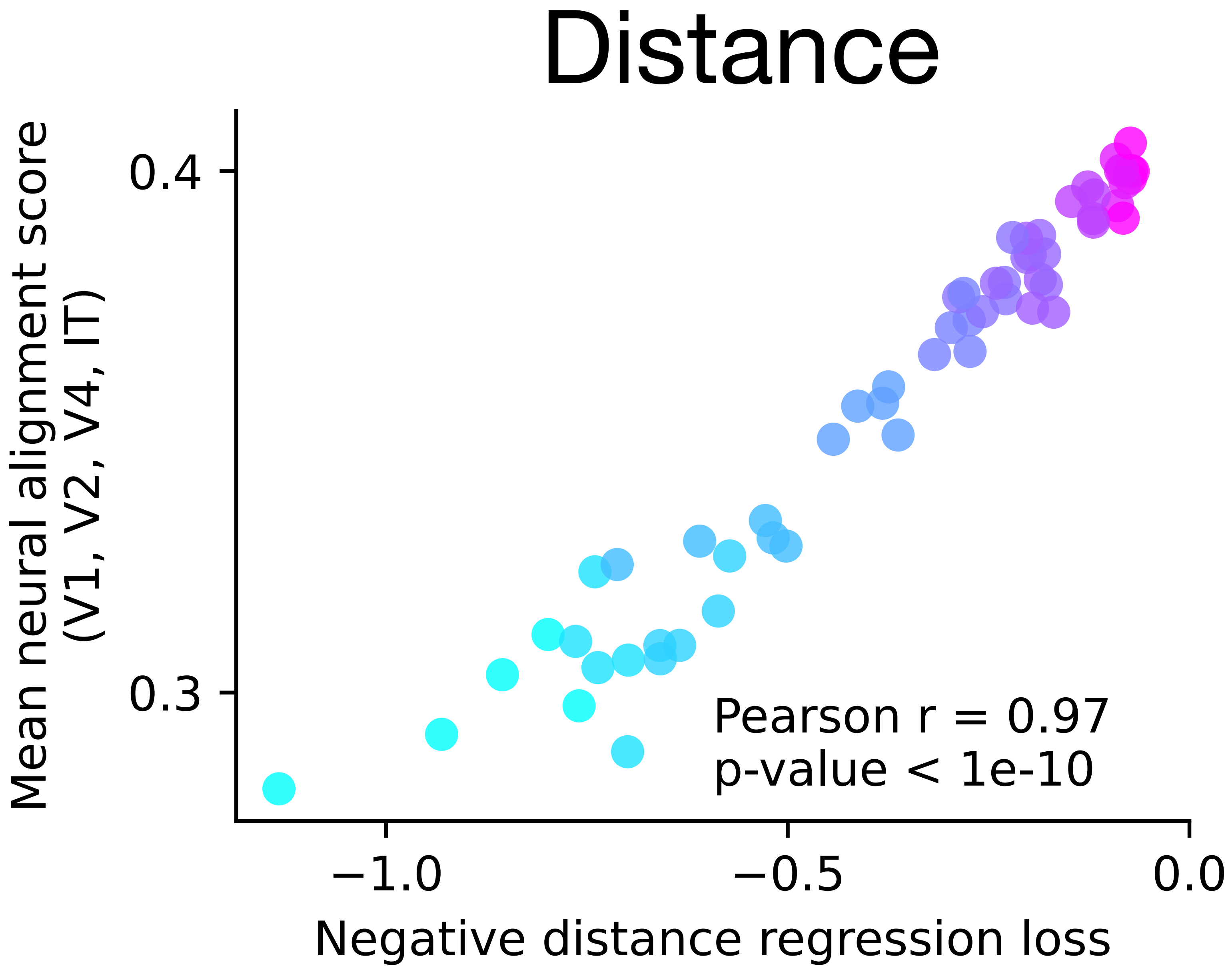
Learning a few spatial latents produces ventral-stream-aligned CNN

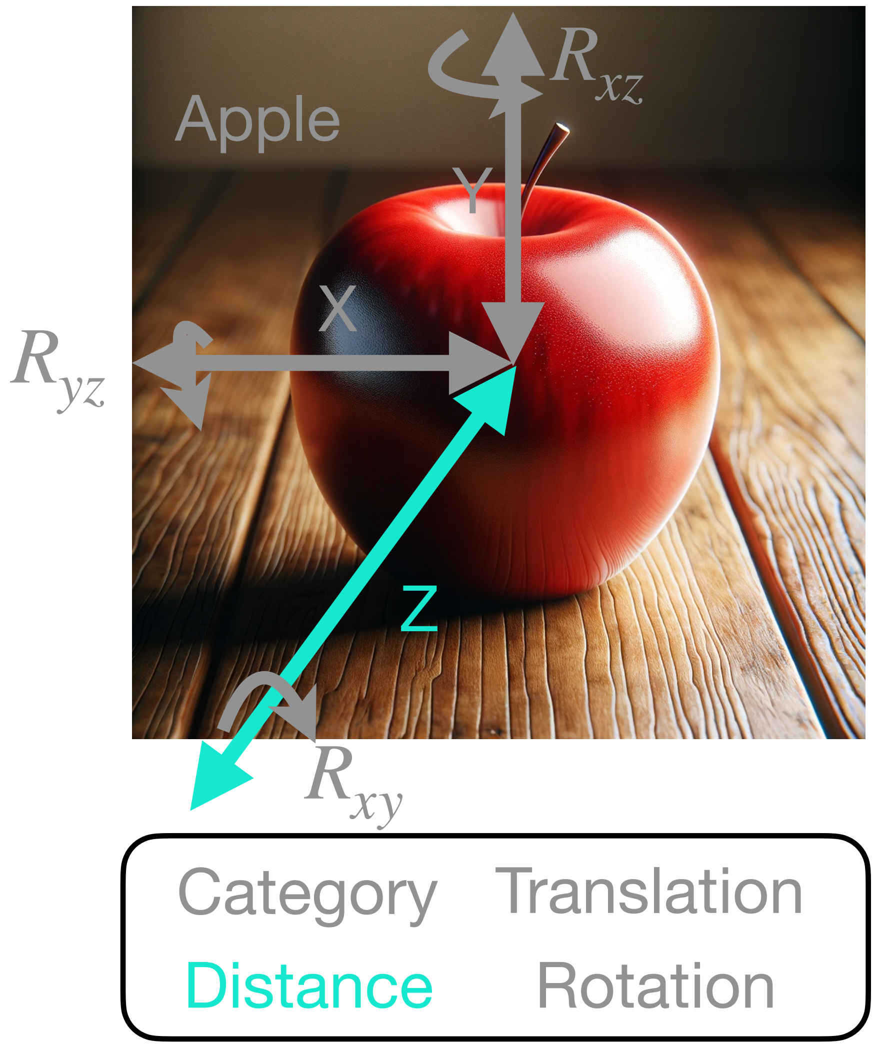


	Training task	# output targets
Spatial latent regression (TDW-117)	Distance regression	1
	Translation regression	2
	Distance + Translation	3
	Rotation regression	6
	Distance + Rotation	7
	Translation + Rotation	8
	Distance + Translation + Rotation	9
Classification (TDW-117)	Object category classification	117
	Object identity classification	548
	All spatial latents + classification	674
Reference	Untrained	NA
	ImageNet-1K classification	1000

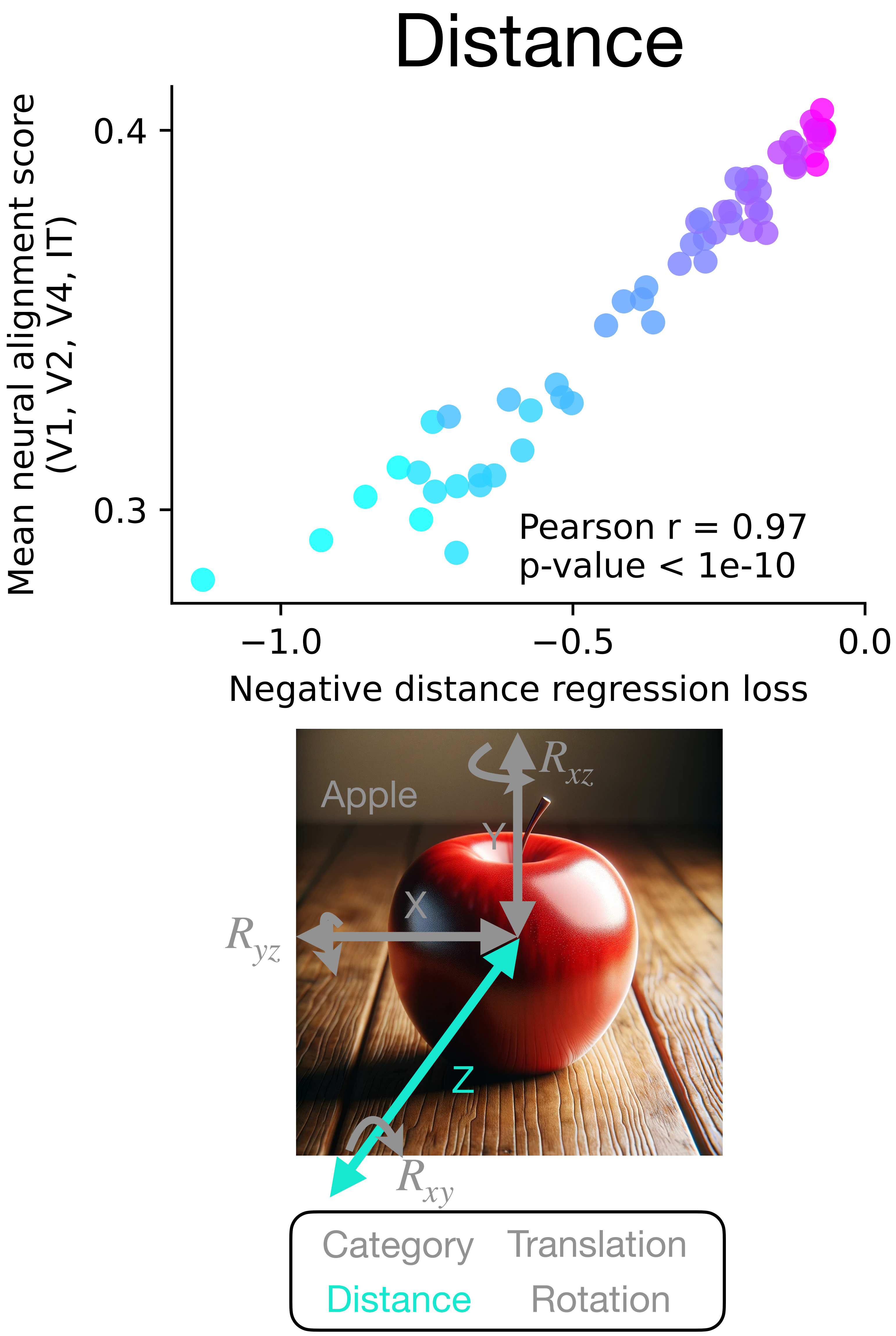
- CNNs trained on a few latents are comparable to those trained on hundreds of categories.

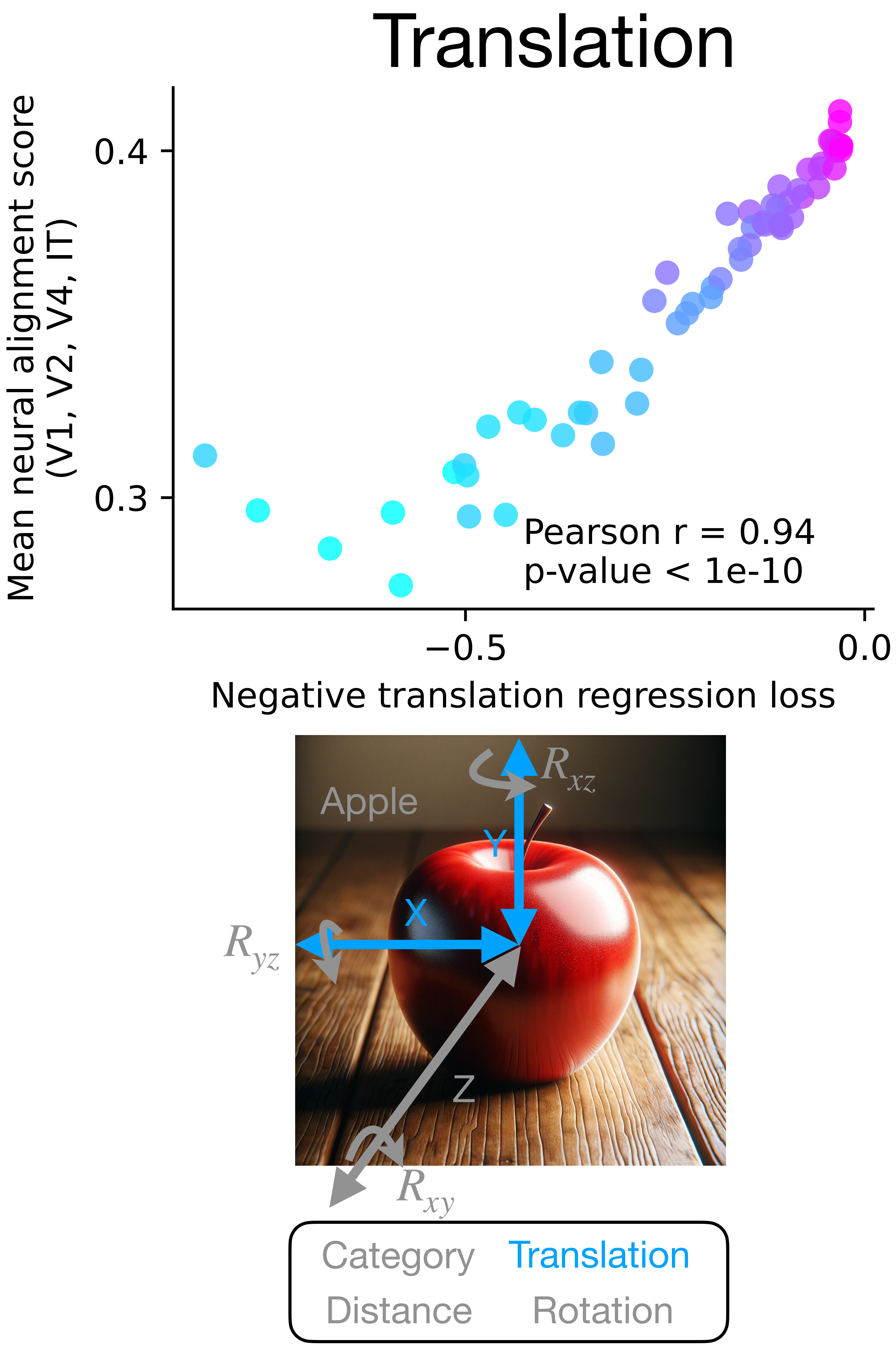
Spatial latents performance also correlates strongly with neural alignment



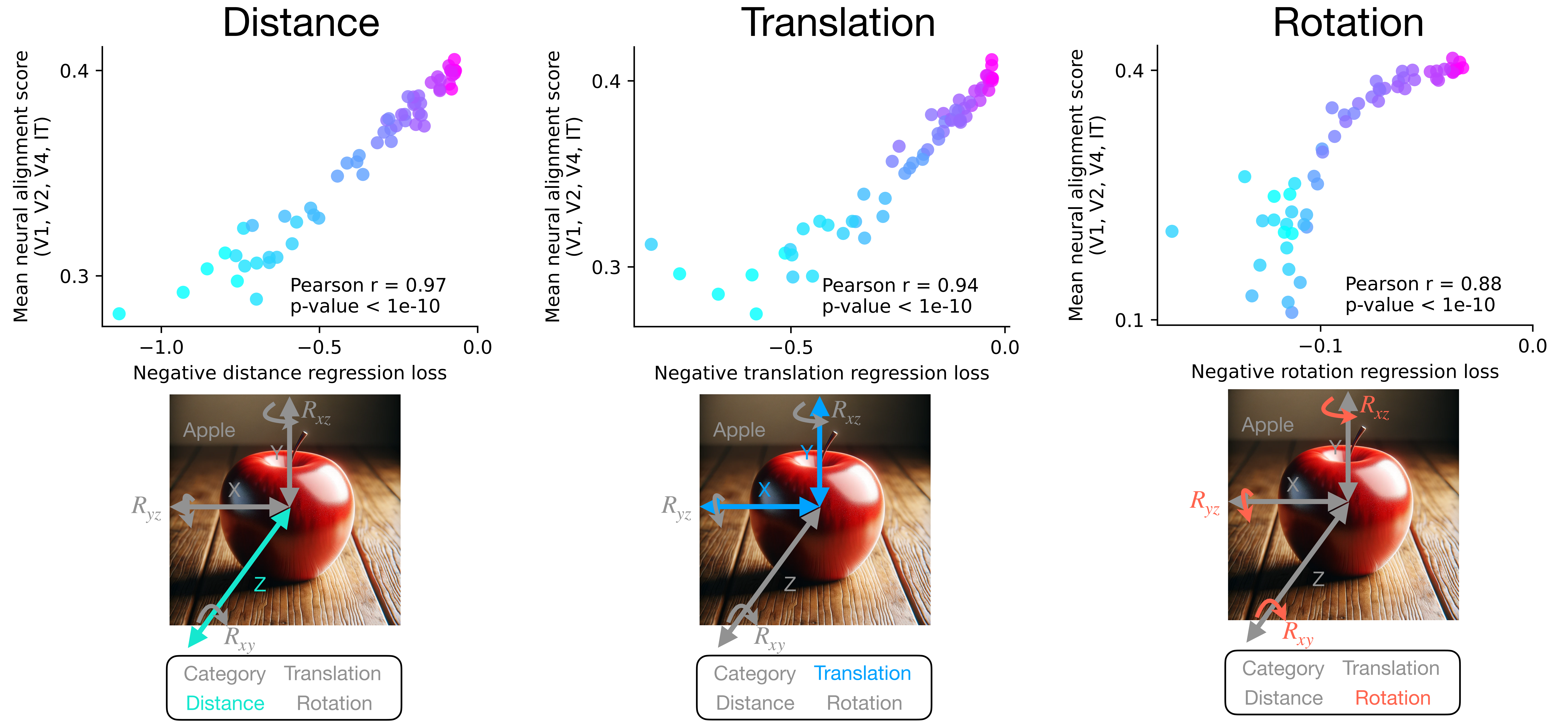


Spatial latents performance also correlates strongly with neural alignment

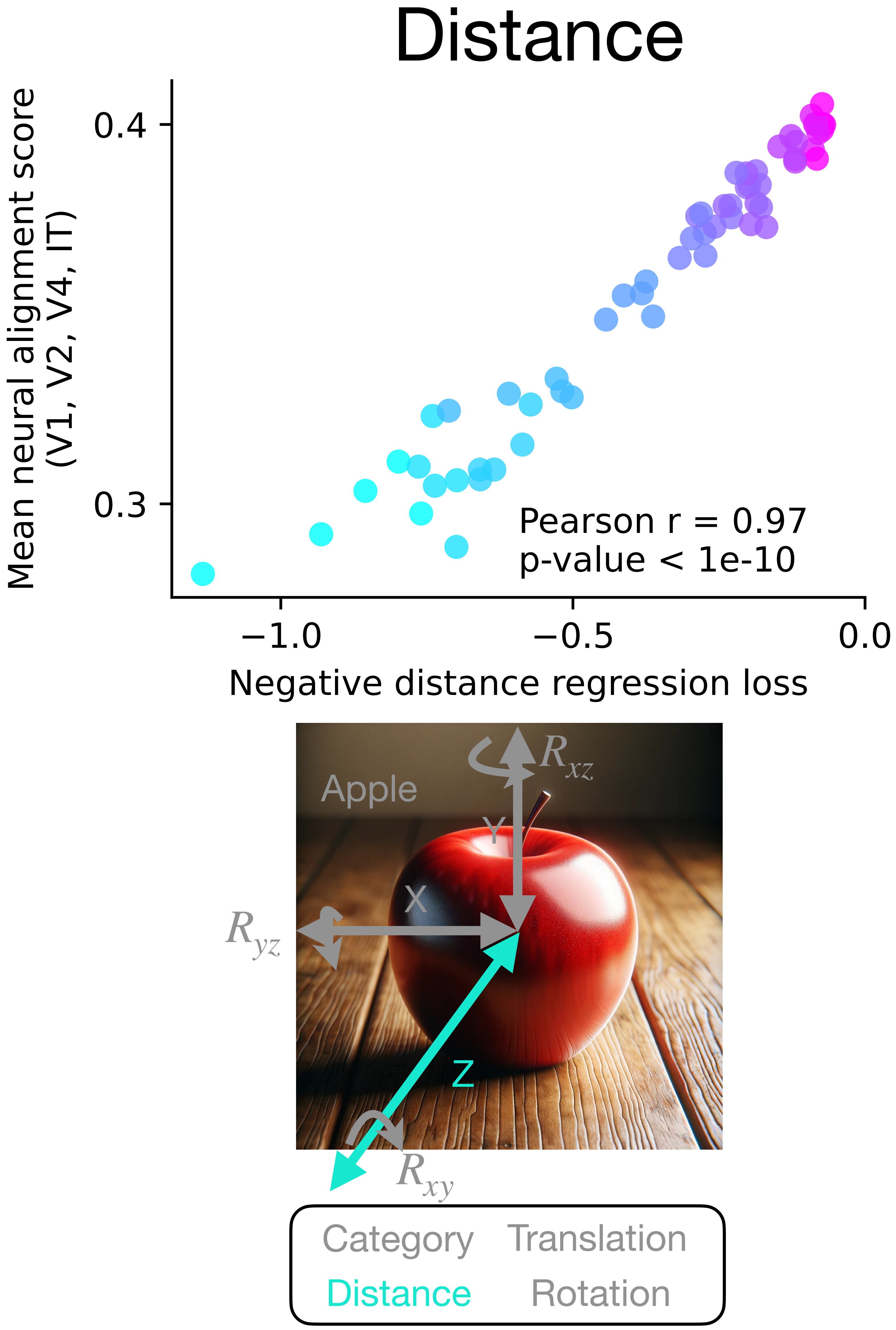


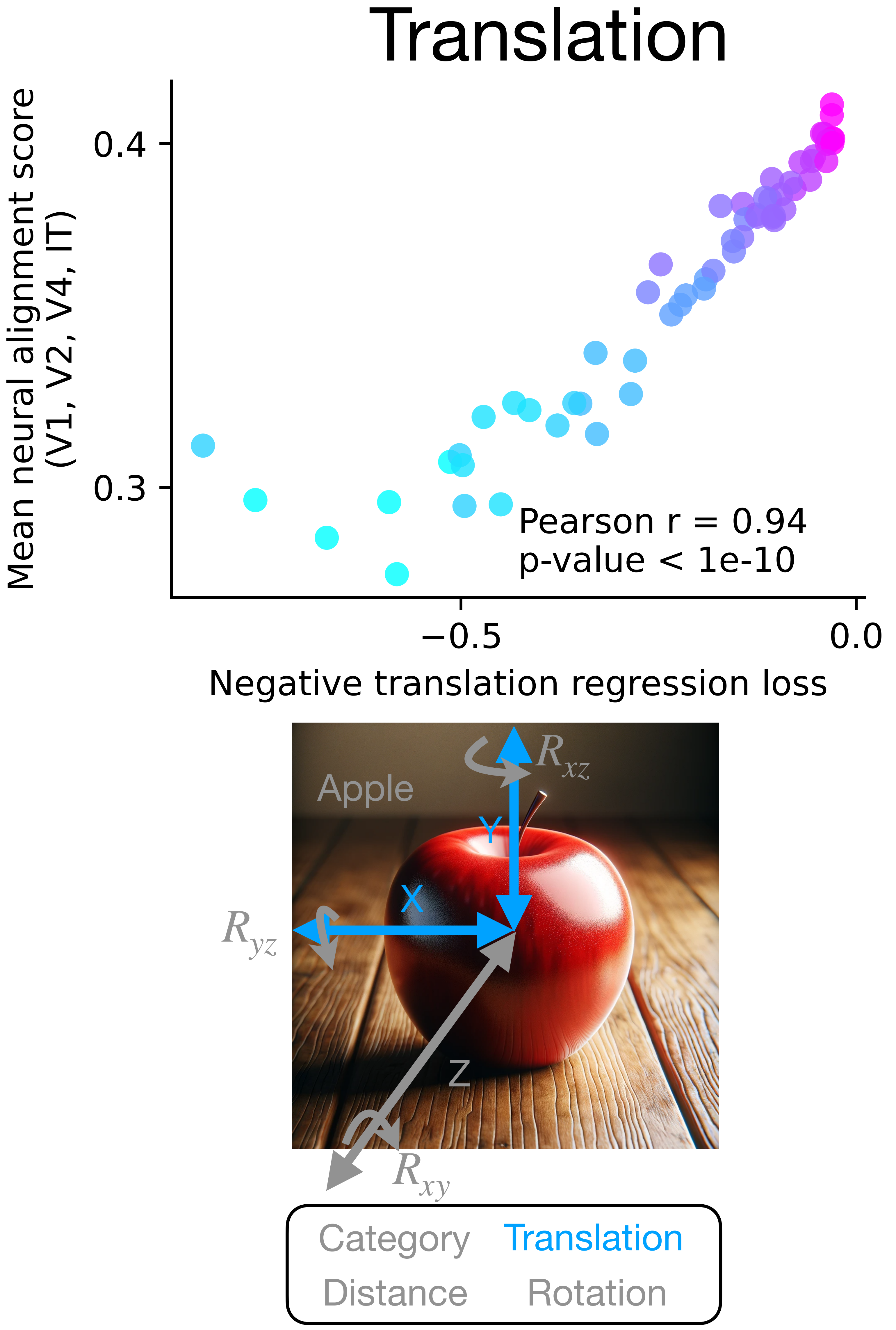


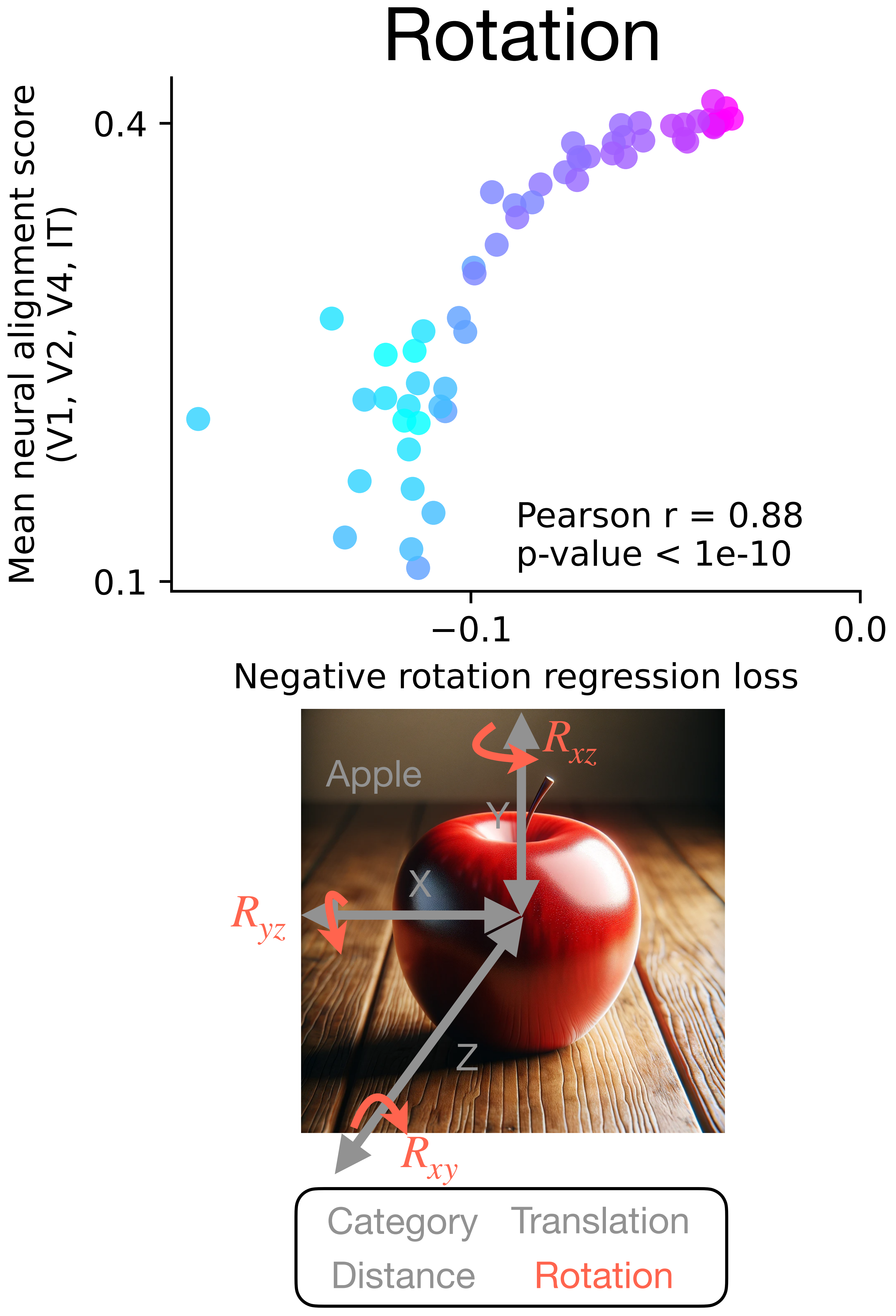
Spatial latents performance also correlates strongly with neural alignment



Spatial latents performance also correlates strongly with neural alignment







- The ventral stream function is also to estimate spatial latents

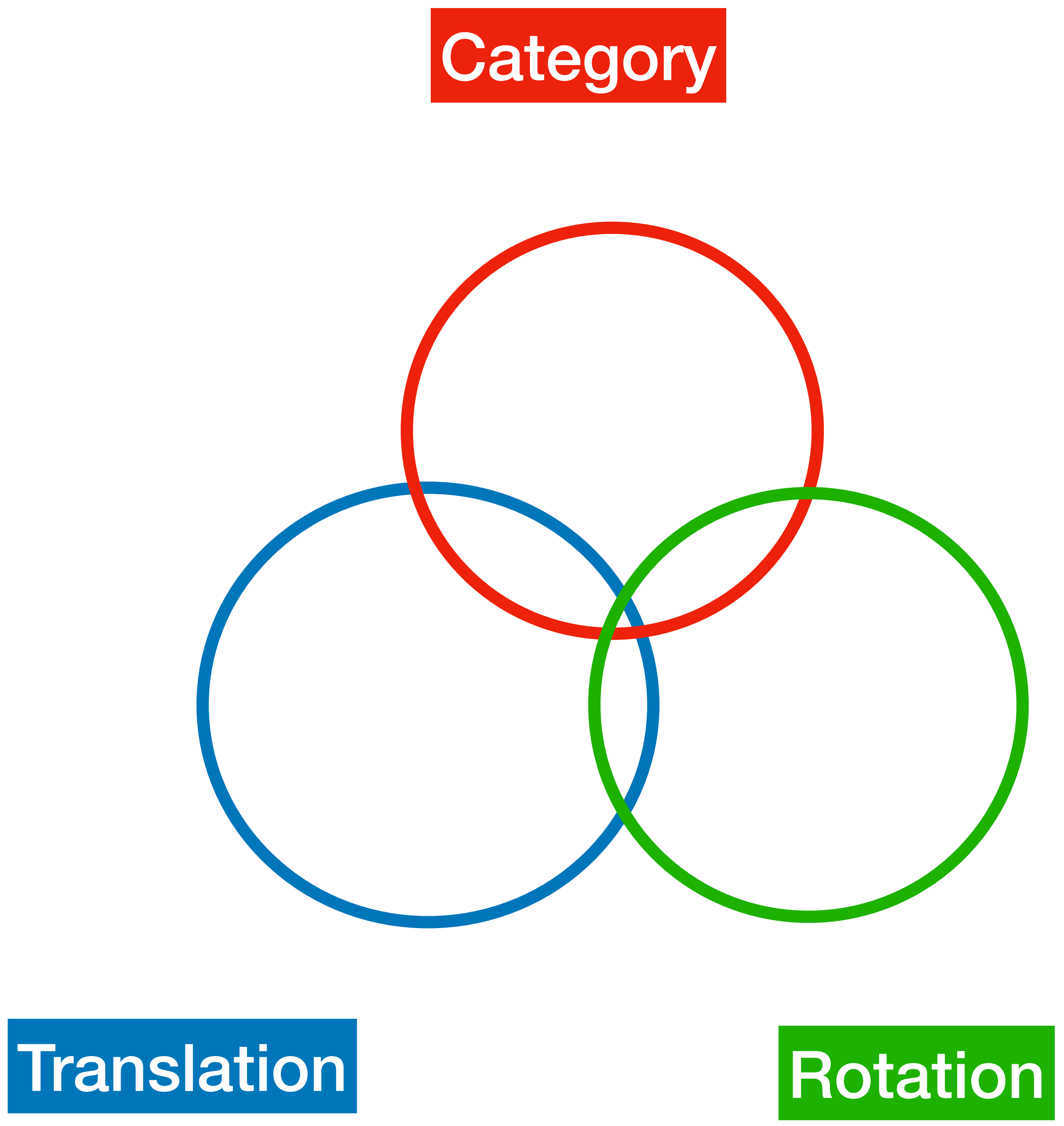
Learning a few spatial latents produced models that has similar neural alignment scores as model trained on categories.

Learning a few spatial latents produced models that has similar neural alignment scores as model trained on categories.

Why is that so? At least two hypotheses:

Learning a few spatial latents produced models that has similar neural alignment scores as model trained on categories.

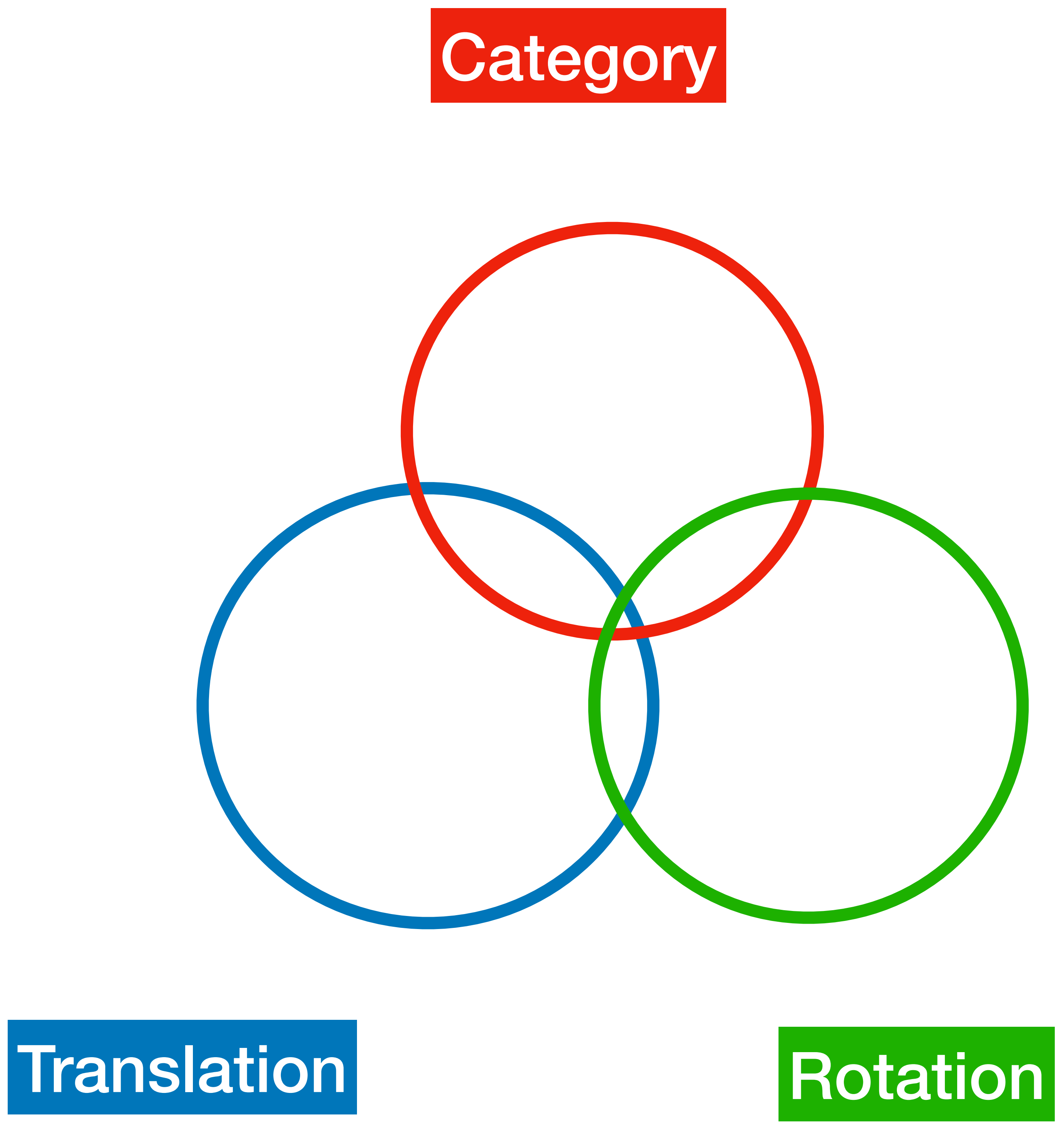
Why is that so? At least two hypotheses:

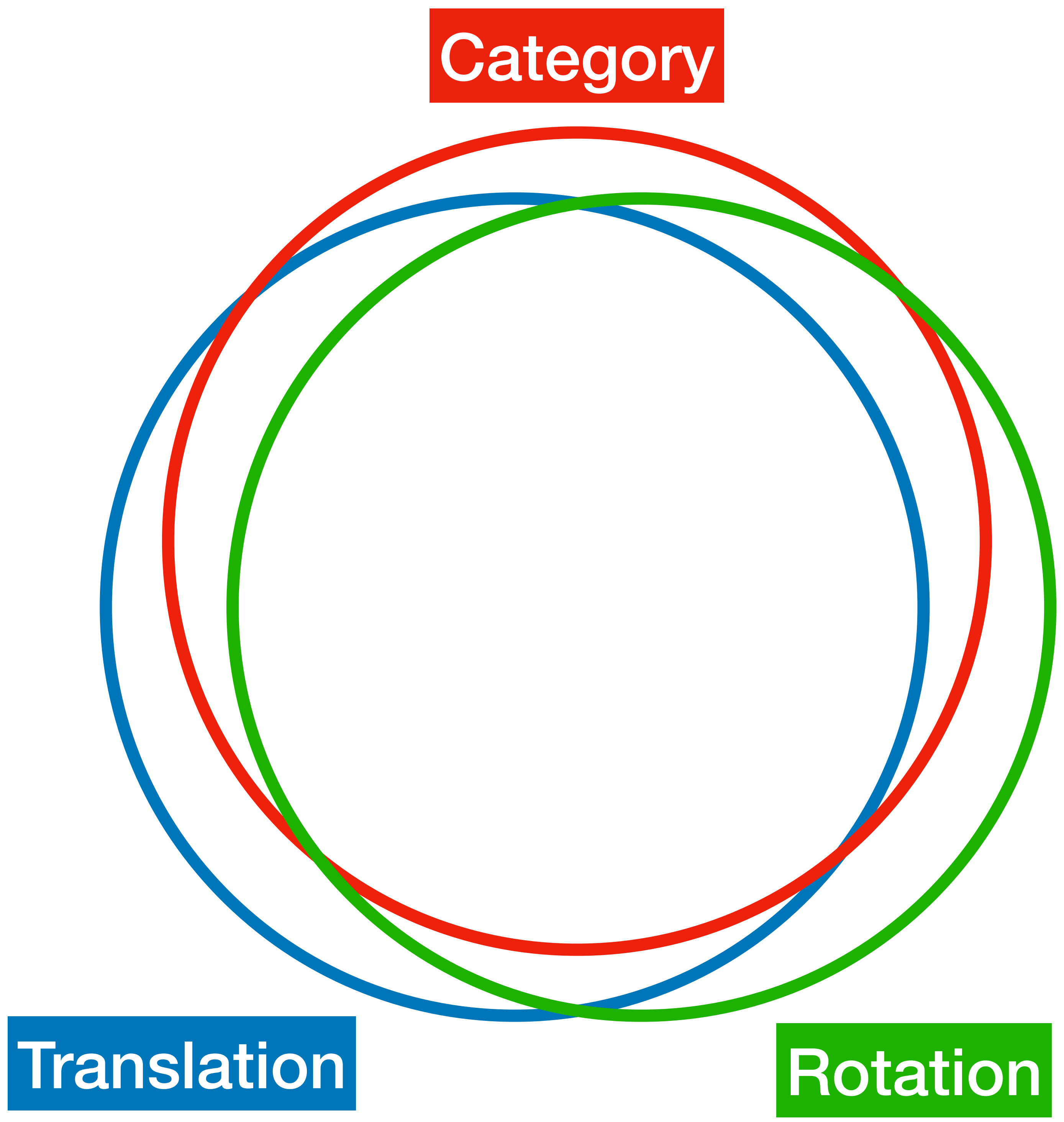


- **Largely dissimilar representations that are equally similar to neural data.**

Learning a few spatial latents produced models that has similar neural alignment scores as model trained on categories.

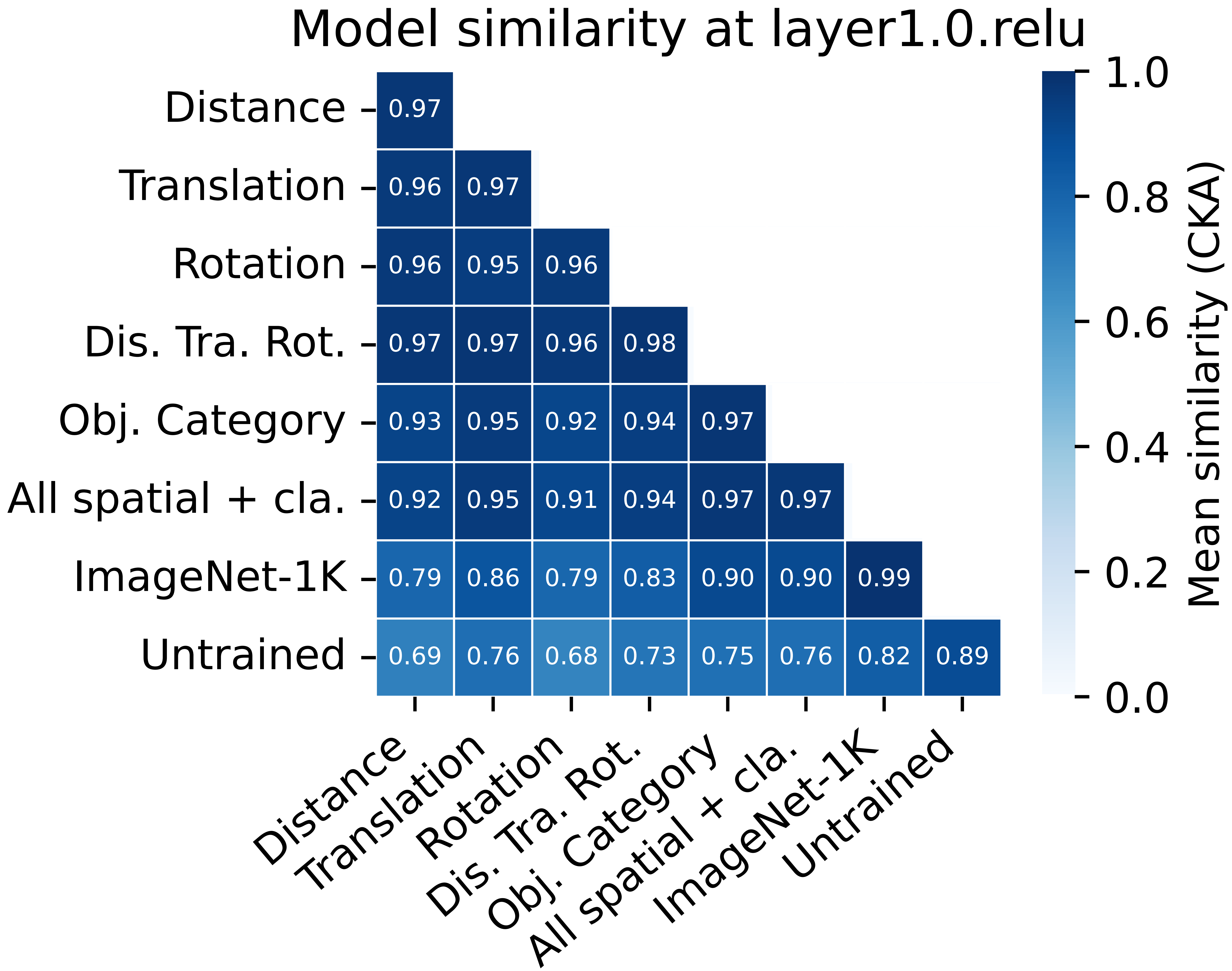
Why is that so? At least two hypotheses:





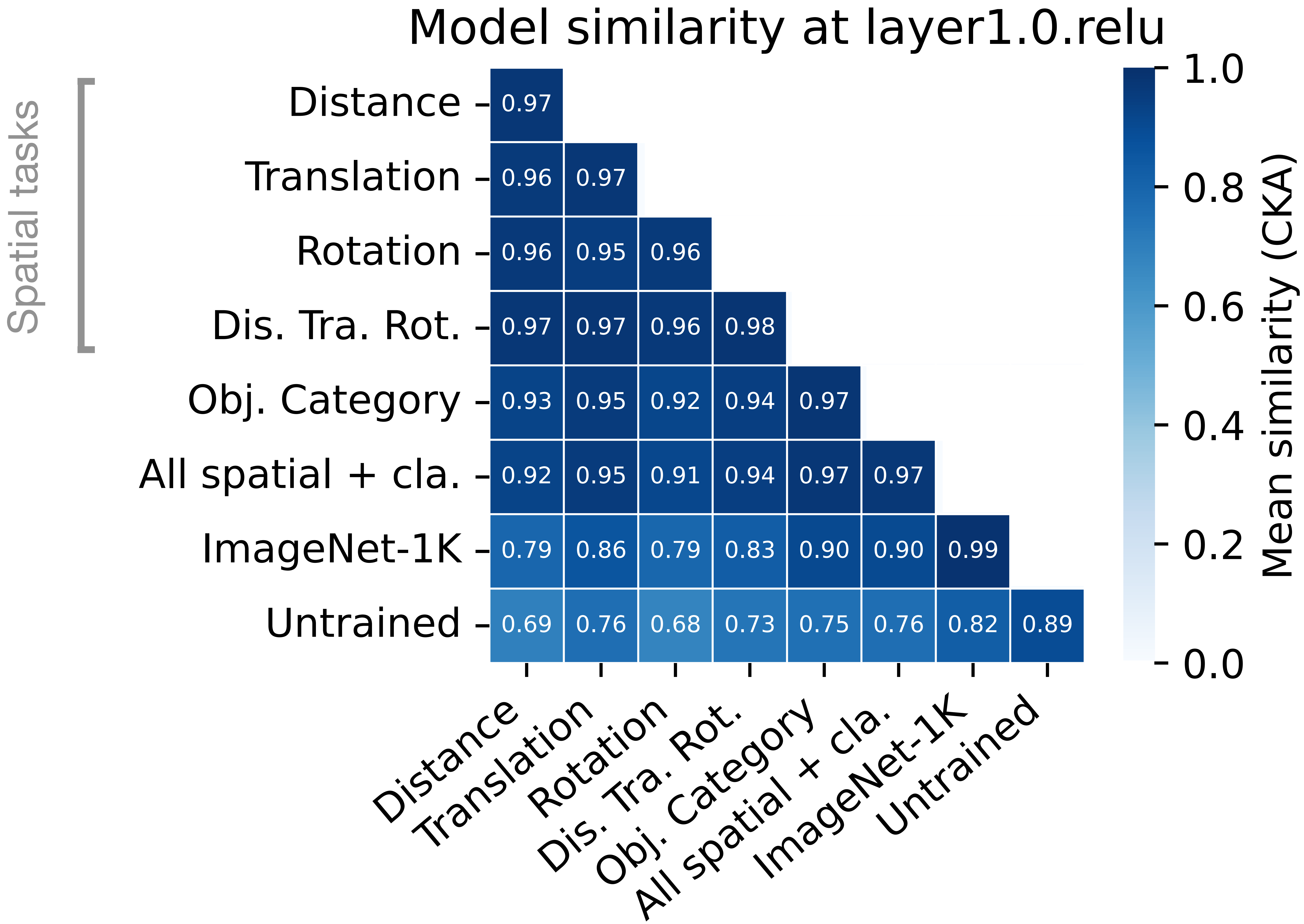
- **Largely dissimilar representations that are equally similar to neural data.**
- **Largely similar representations, thus, similar alignment scores.**

Different objectives lead to similar representations in early layers



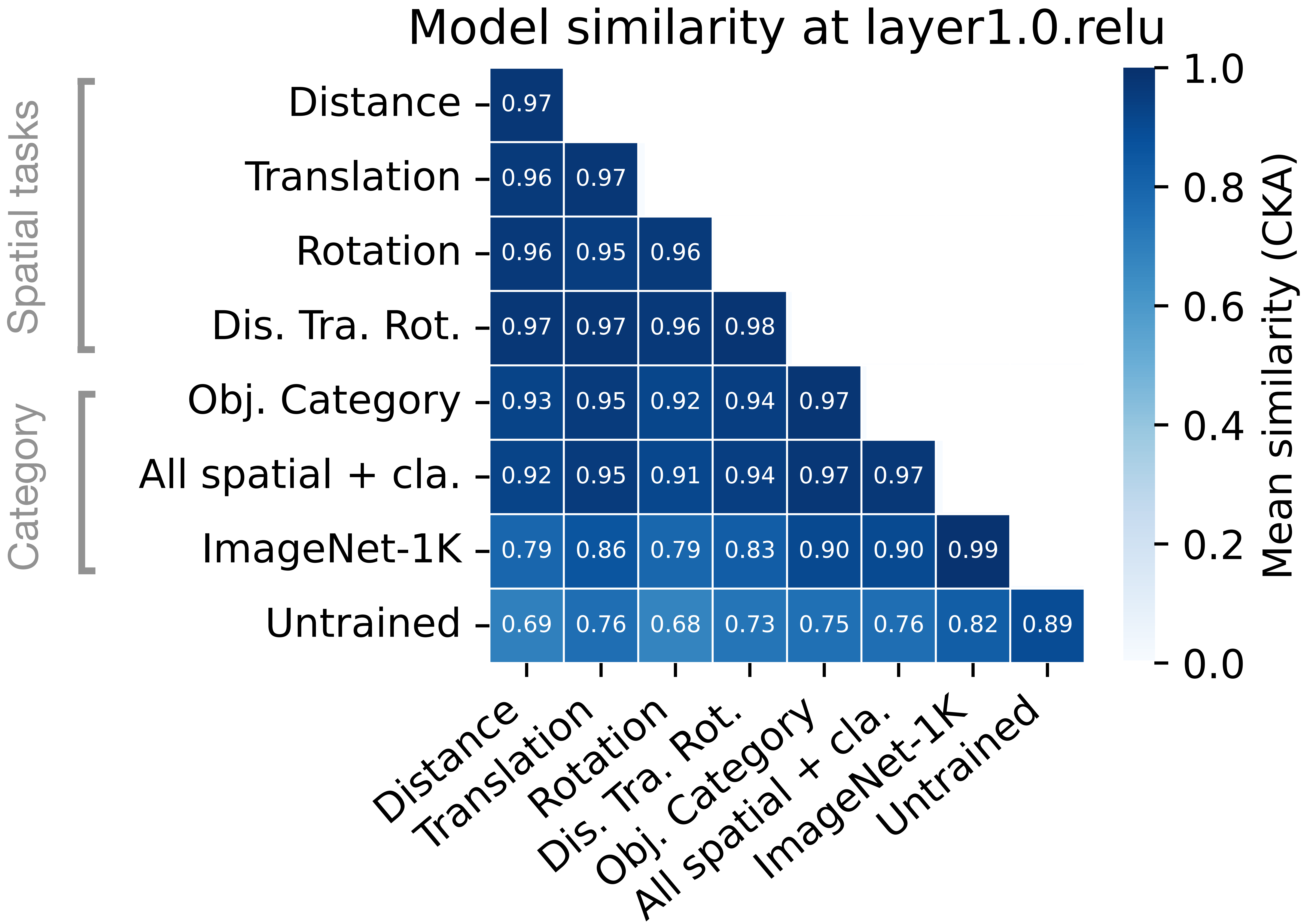
- CKA analysis revealed models learned very similar representations despite trained on very different tasks.

Different objectives lead to similar representations in early layers



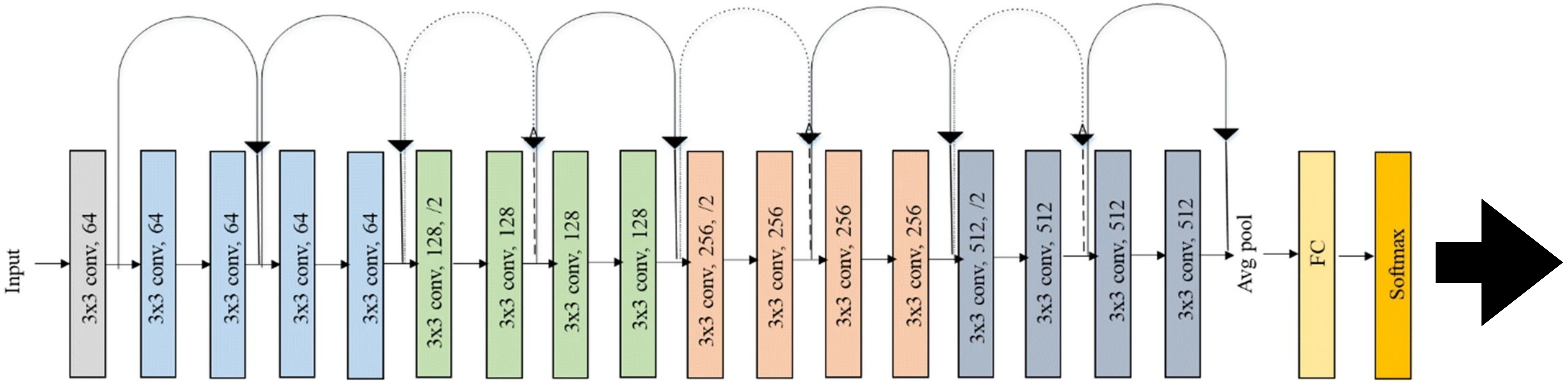
- CKA analysis revealed models learned very similar representations despite trained on very different tasks.

Different objectives lead to similar representations in early layers



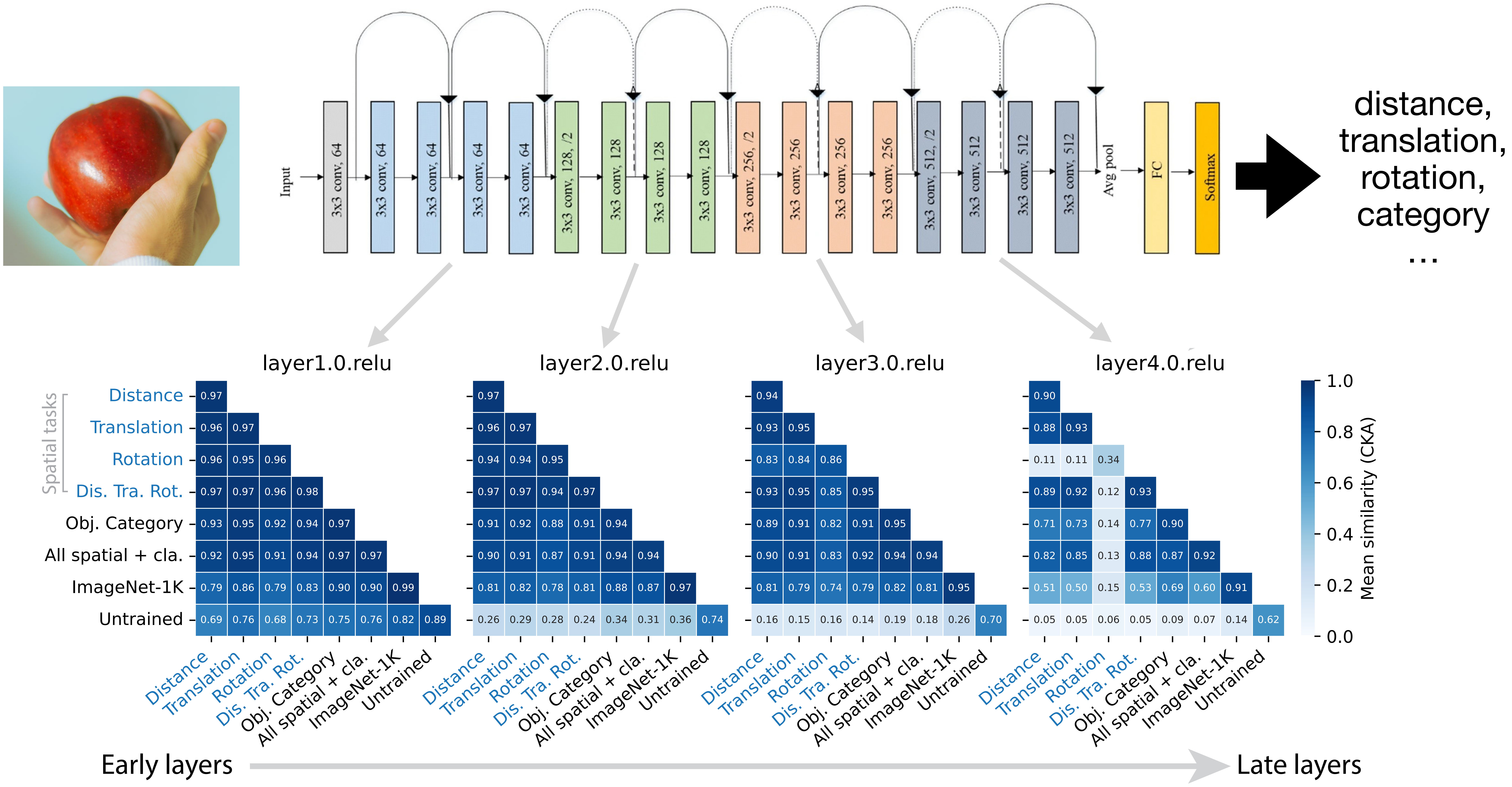
- CKA analysis revealed models learned very similar representations despite trained on very different tasks.

Representations are similar in early to middle, but diverge at late layers

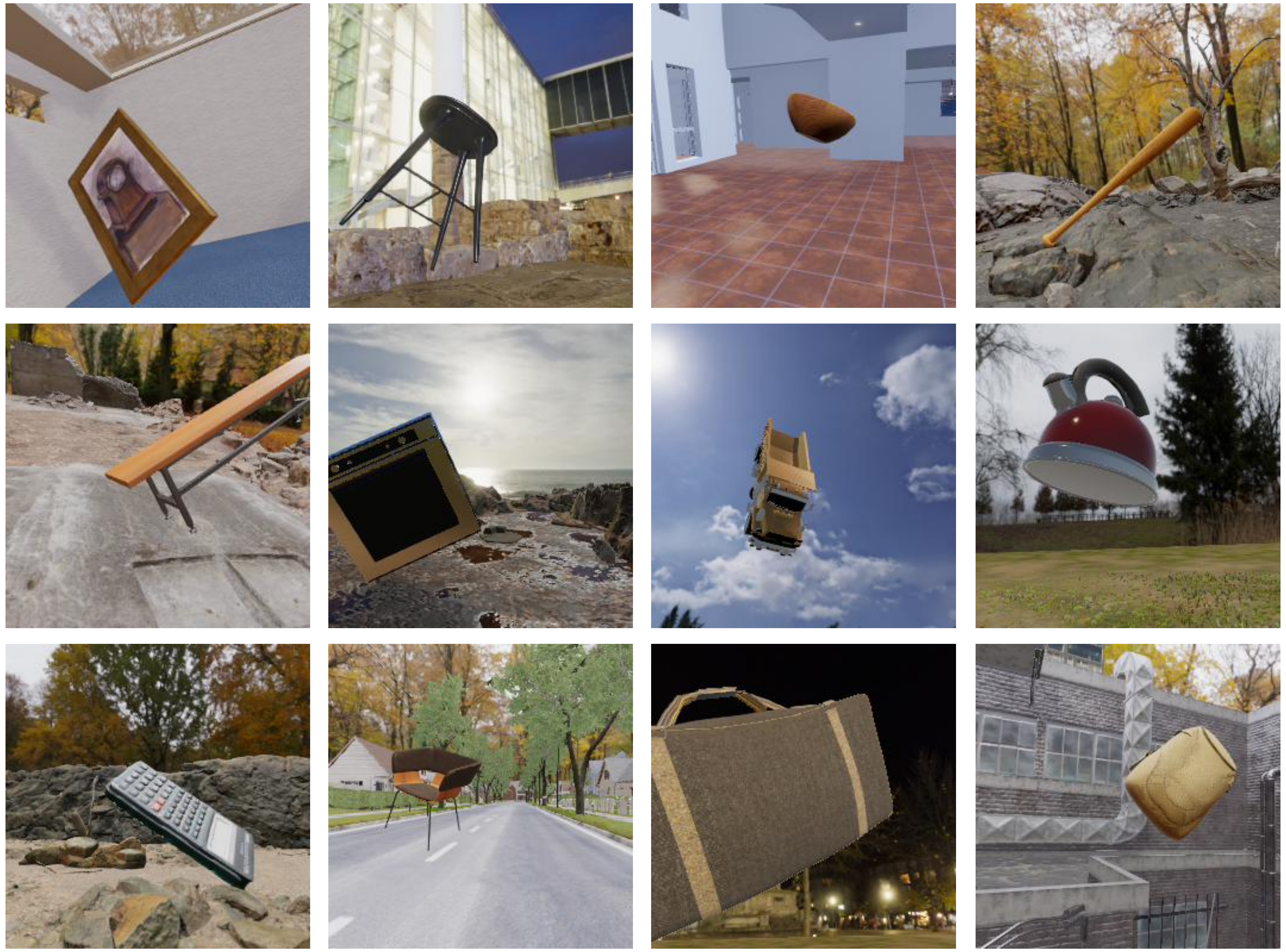


distance,
translation,
rotation,
category
...

Representations are similar in early to middle, but diverge at late layers

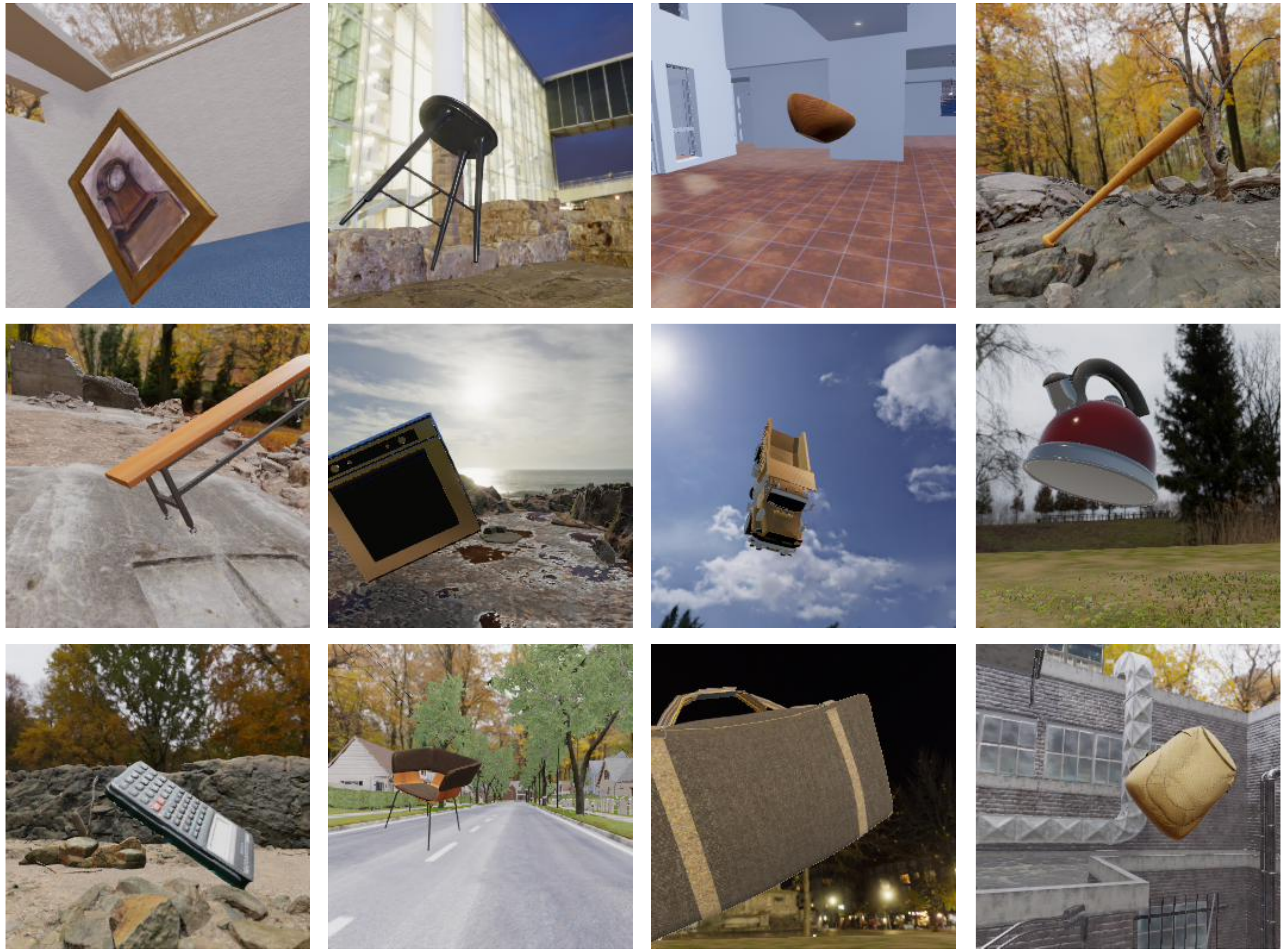


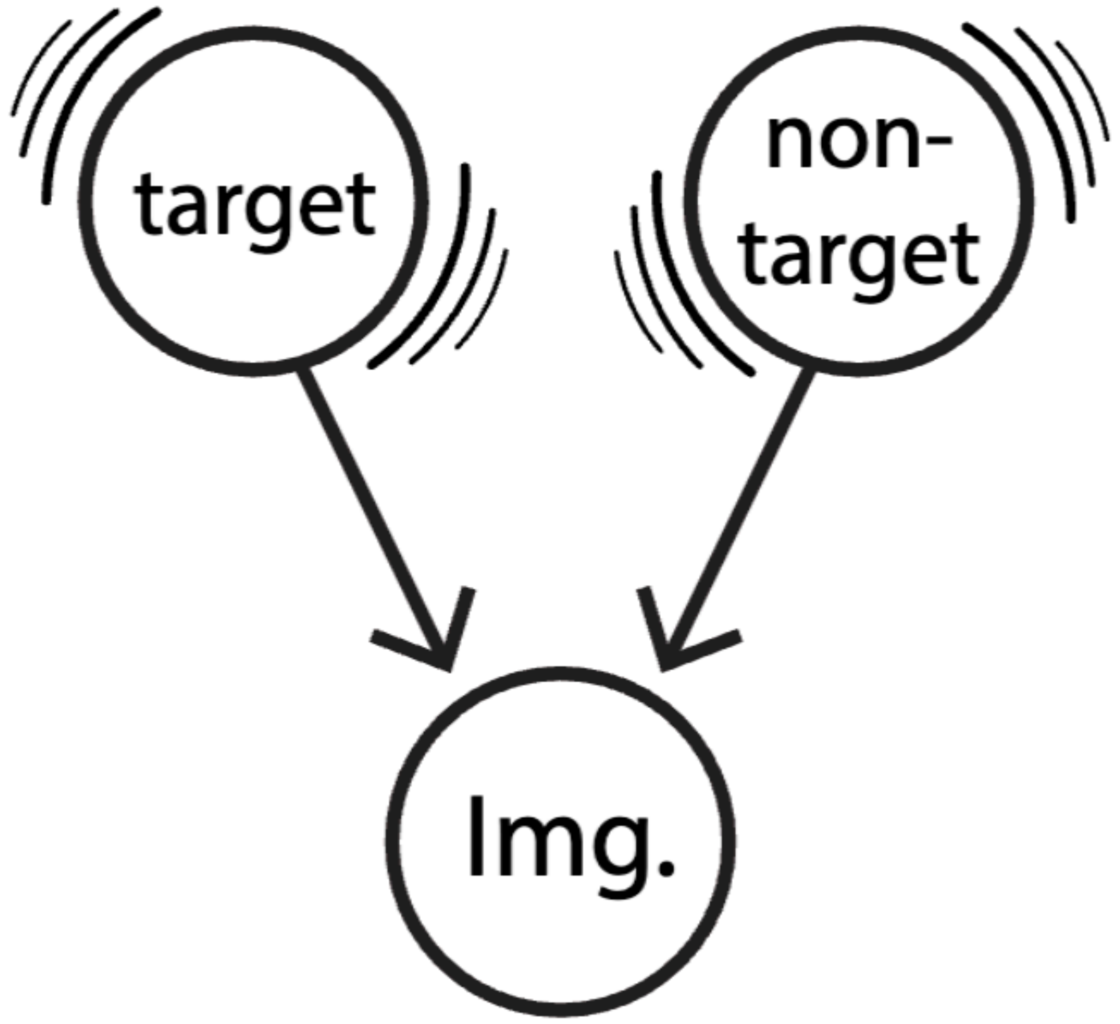
Why do models learn similar representations?
(Despite being trained on very different tasks)



- Models are trained on the same dataset although different tasks.

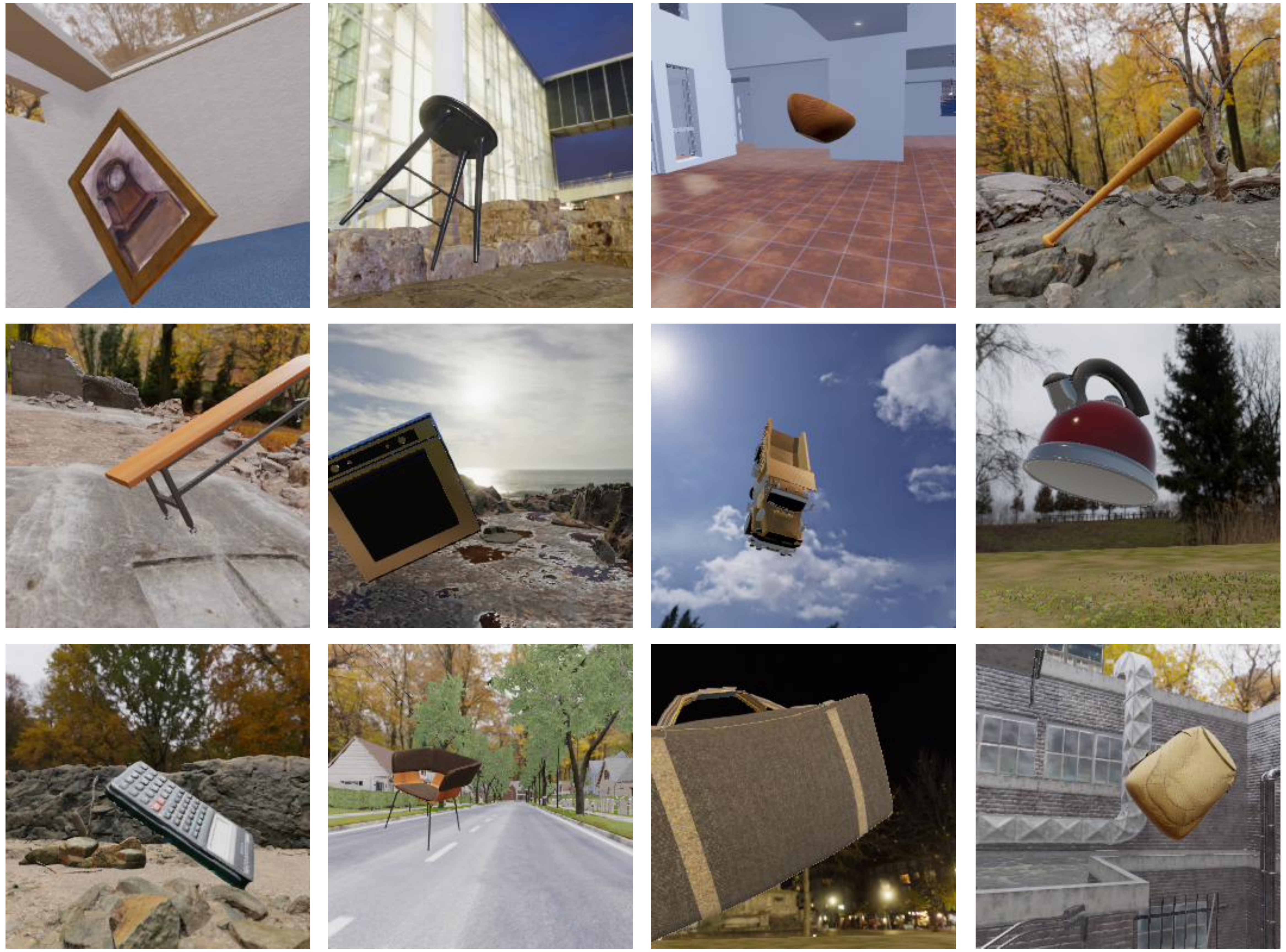
TDW ThreeDWorld





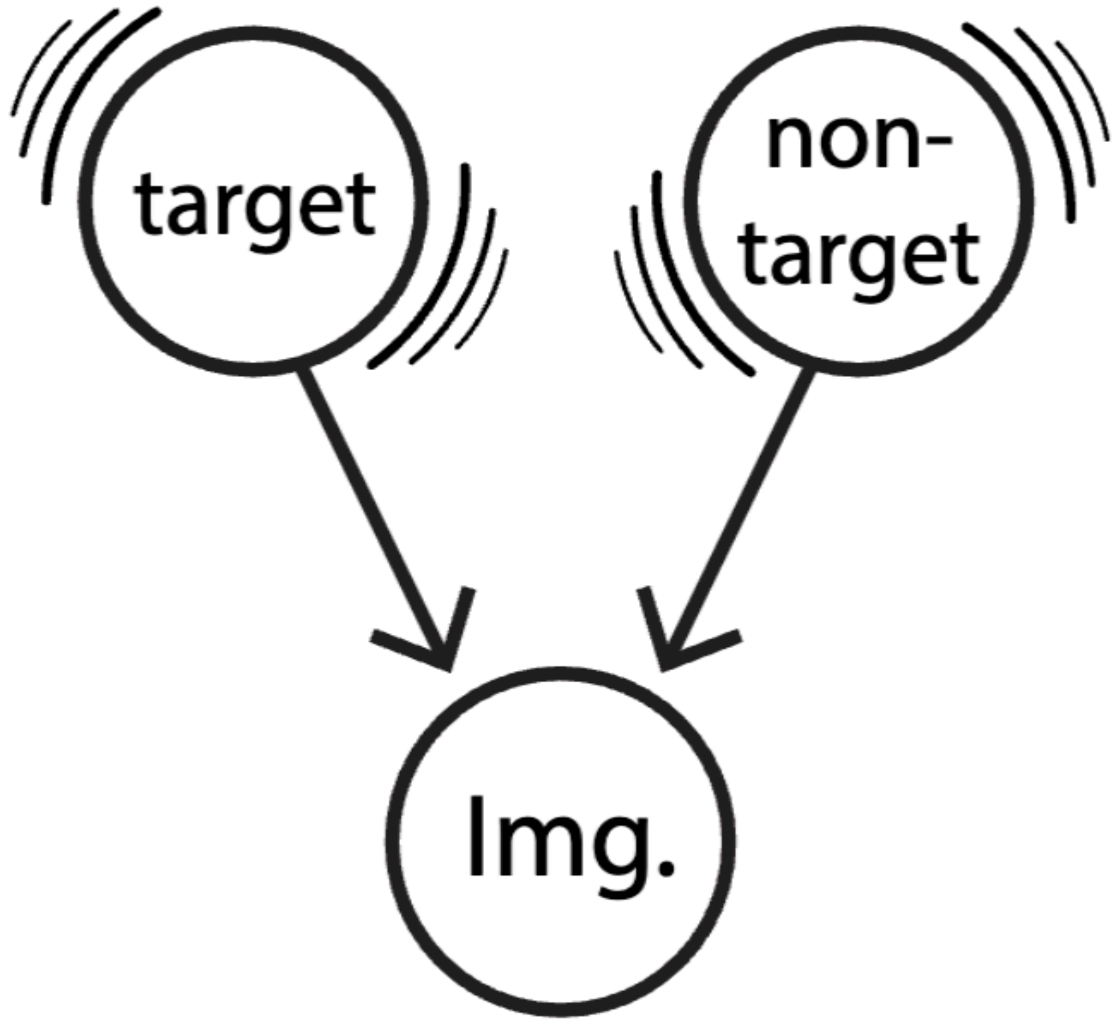
- Models are trained on the same dataset although different tasks.

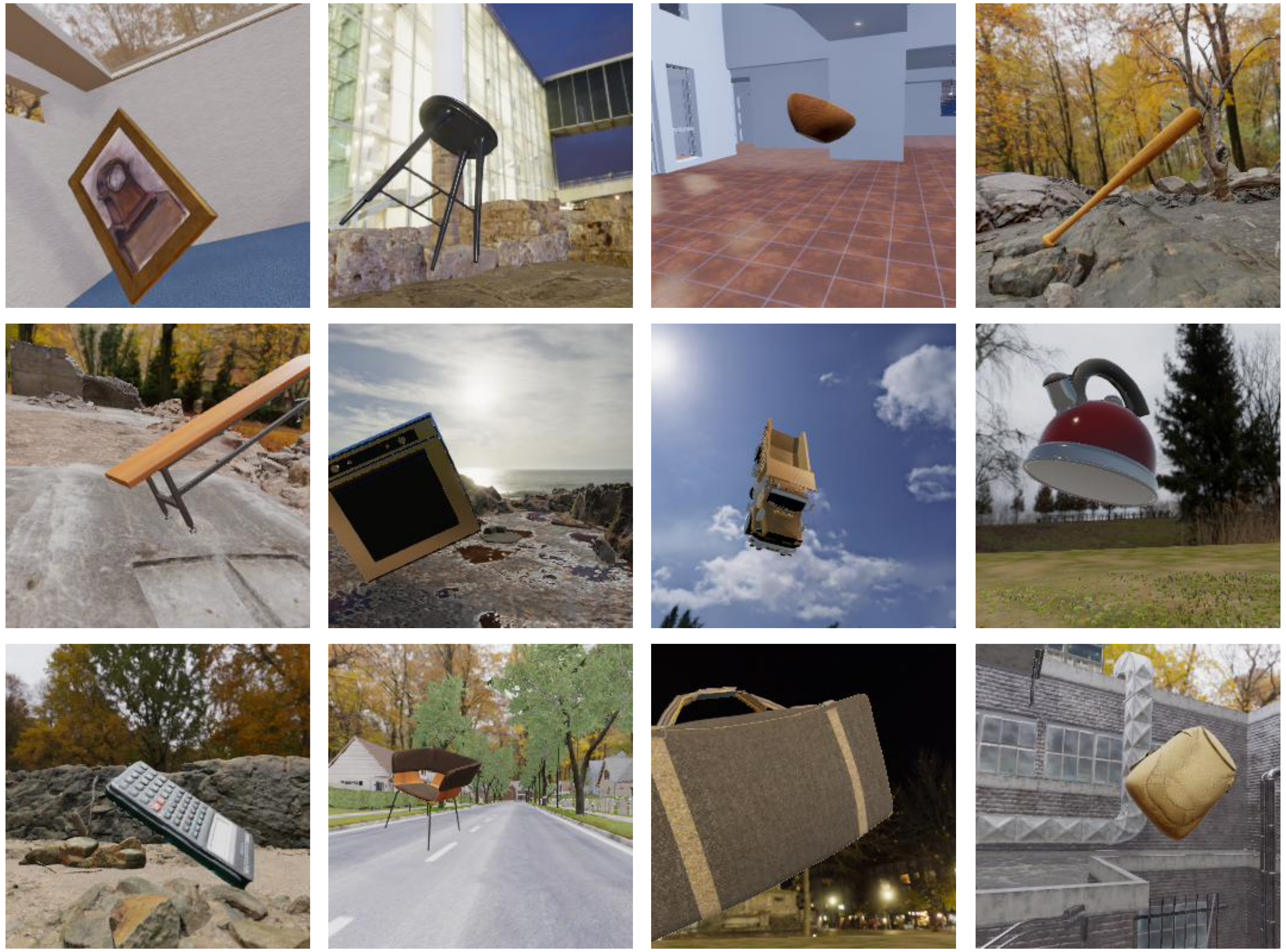
TDW ThreeDWorld



- Models are trained on the same dataset although different tasks.

Our hypothesis:

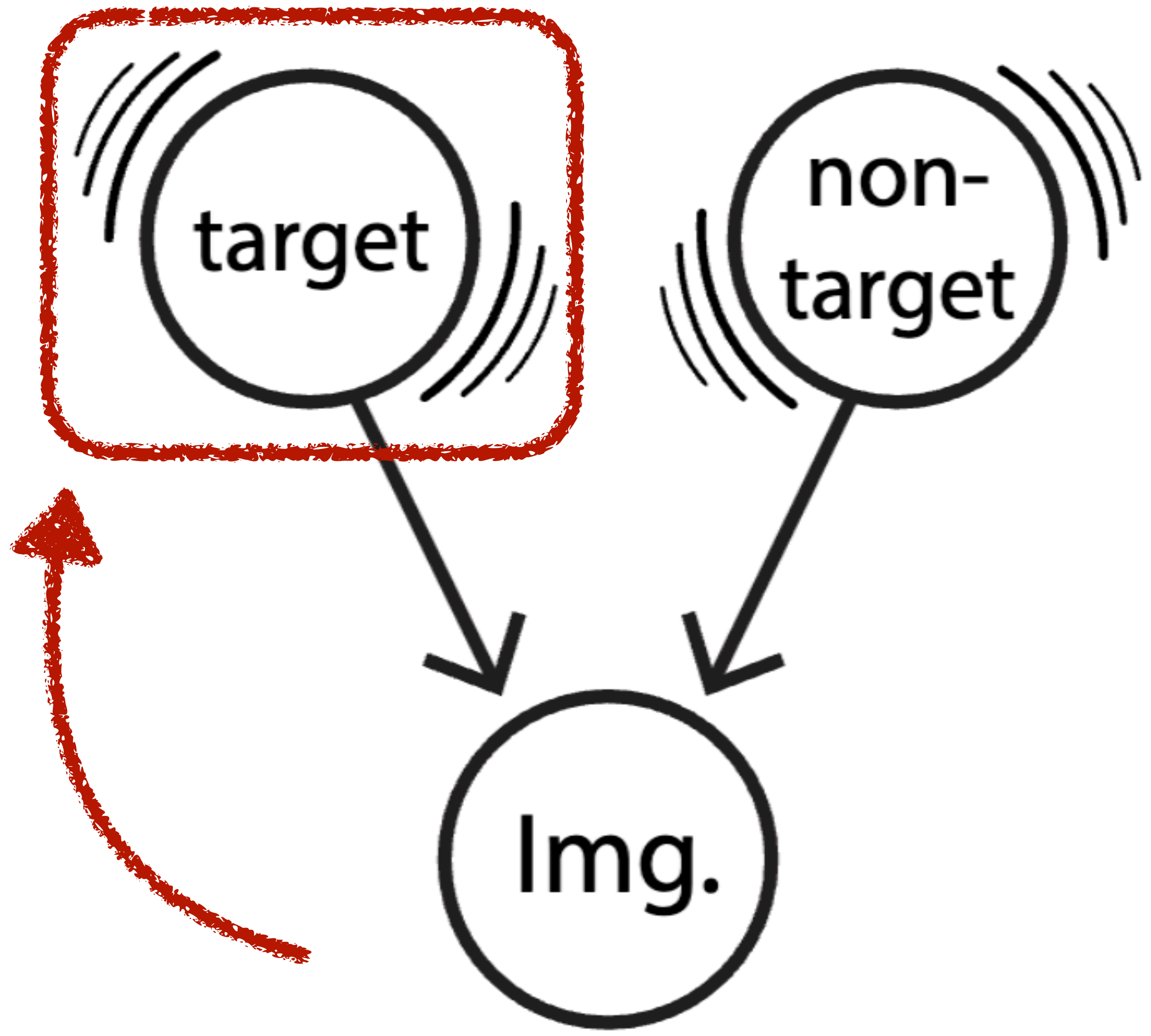


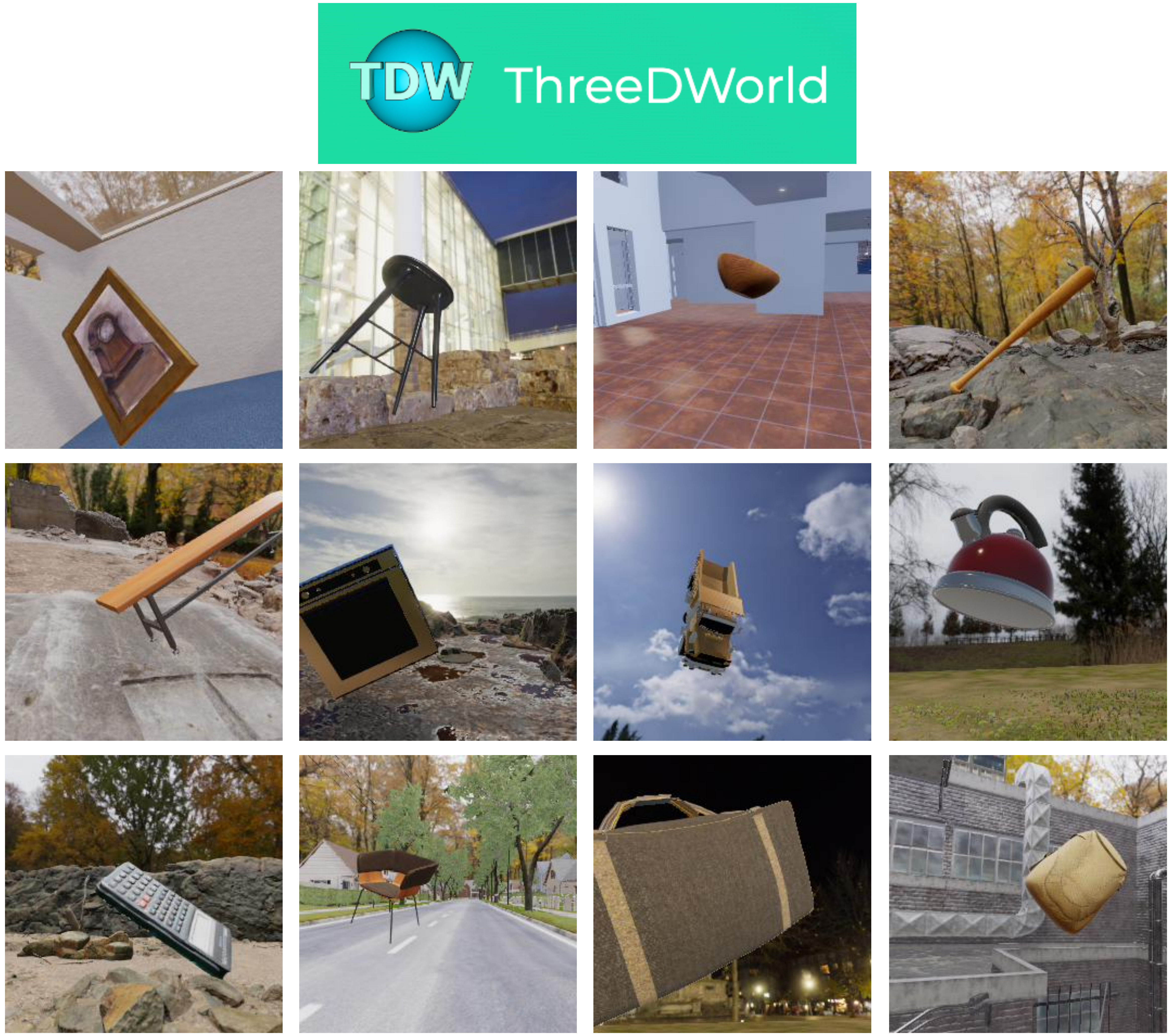


- Models are trained on the same dataset although different tasks.

Our hypothesis:

- learn the target latent

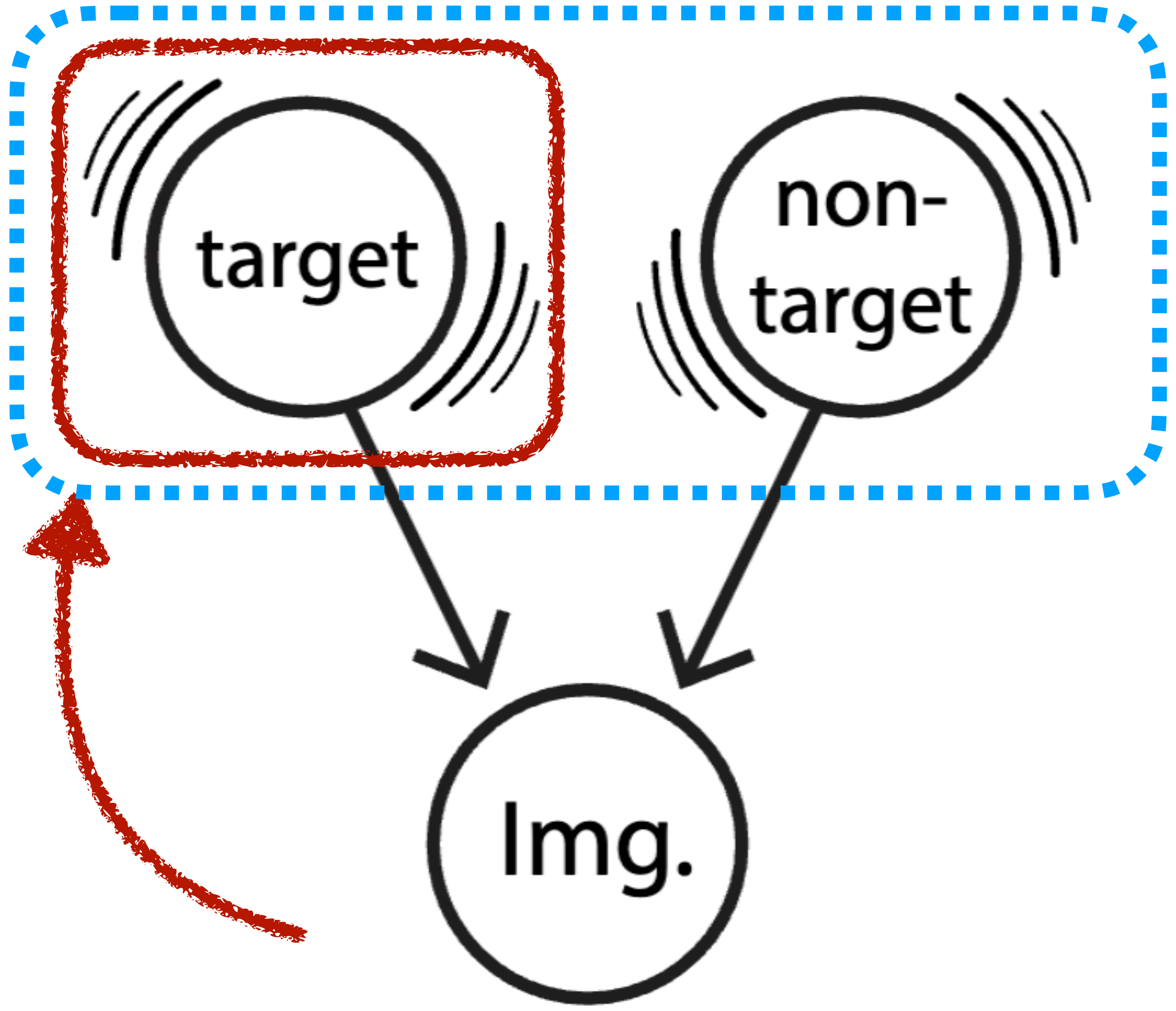




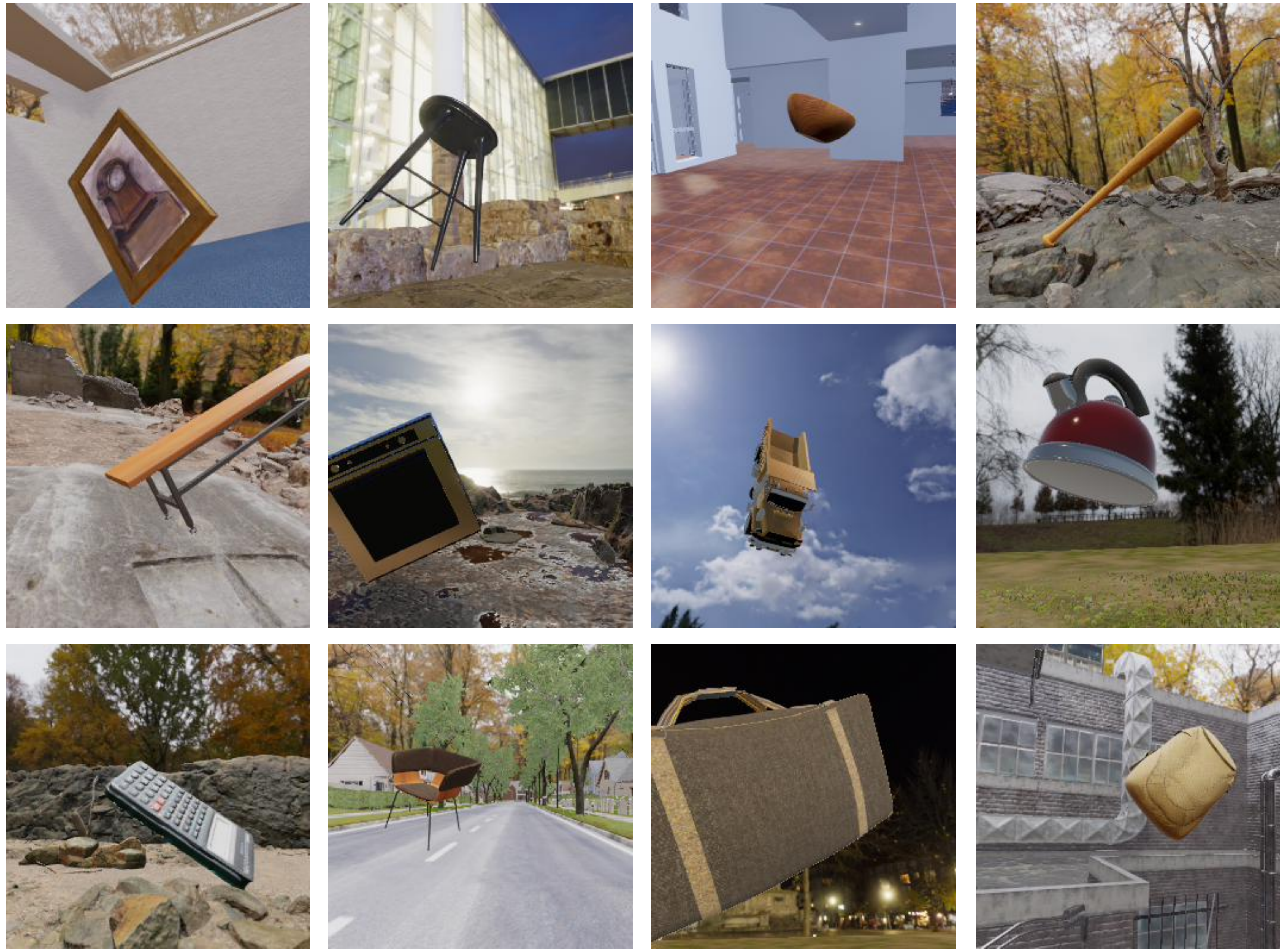
- Models are trained on the same dataset although different tasks.

Our hypothesis:

- learn the target latent
- **inadvertently** learned non-target latents due to their variability.

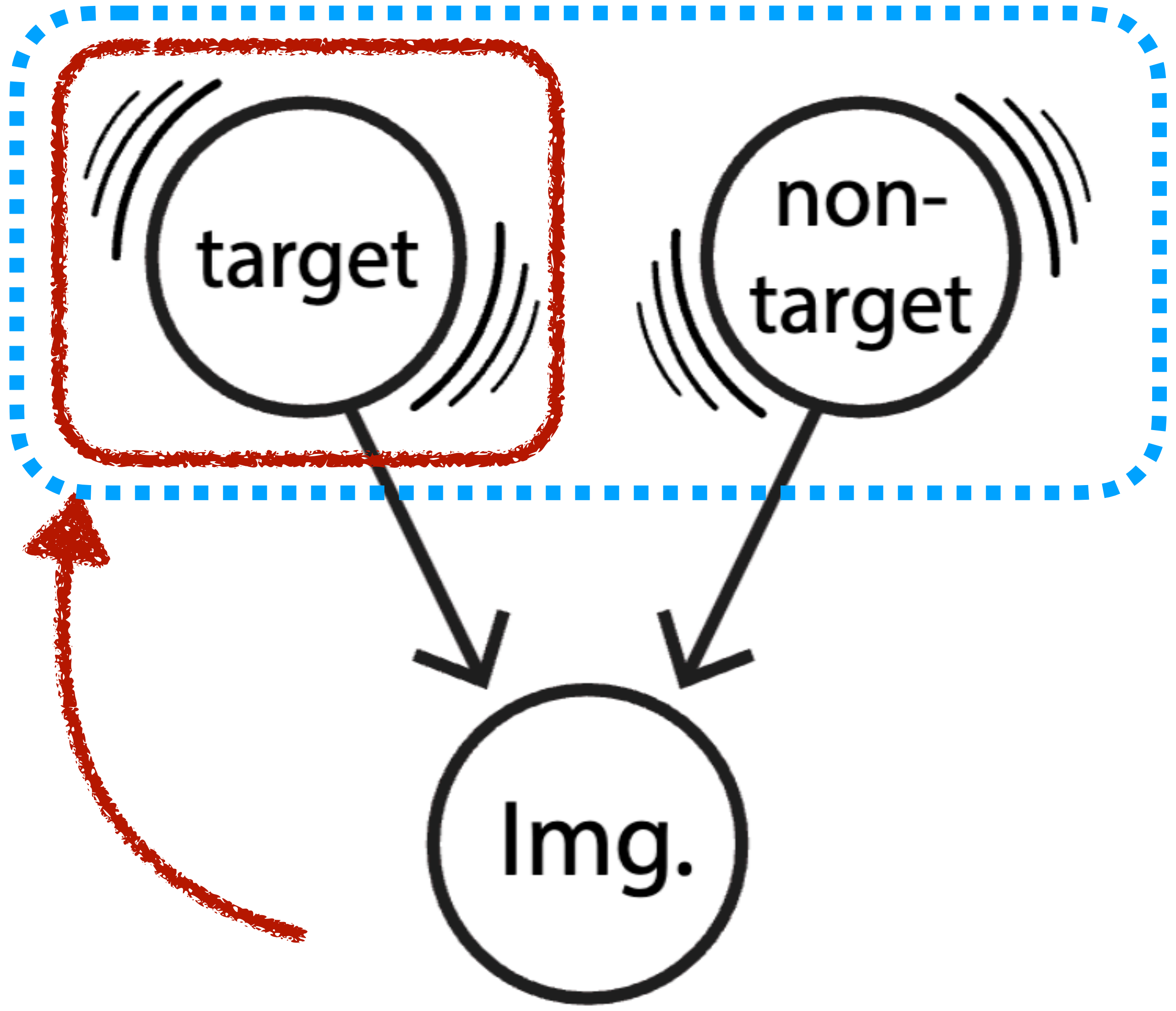


TDW ThreeDWorld



Our hypothesis:

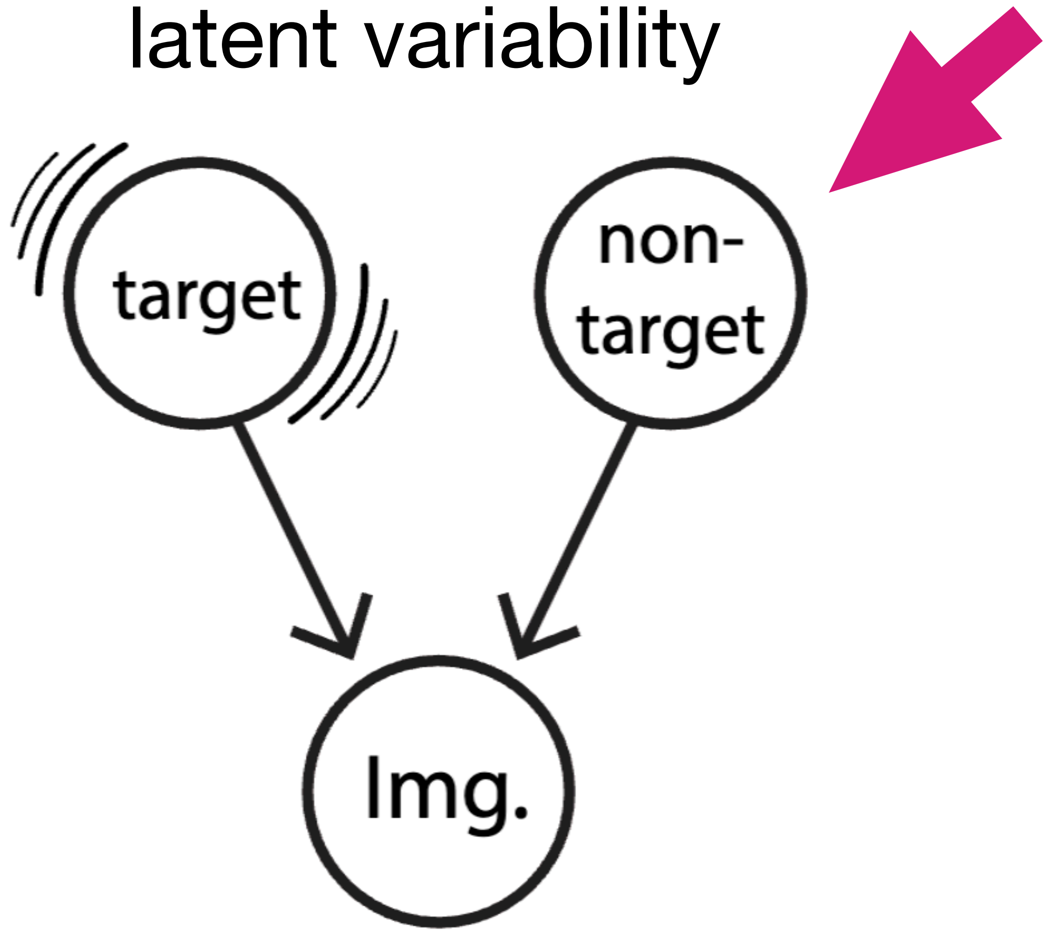
- learn the target latent
- **inadvertently** learned non-target latents due to their variability.



- Models are trained on the same dataset although different tasks.
- As a model learns a single task, it simultaneously learns the structure of the world.

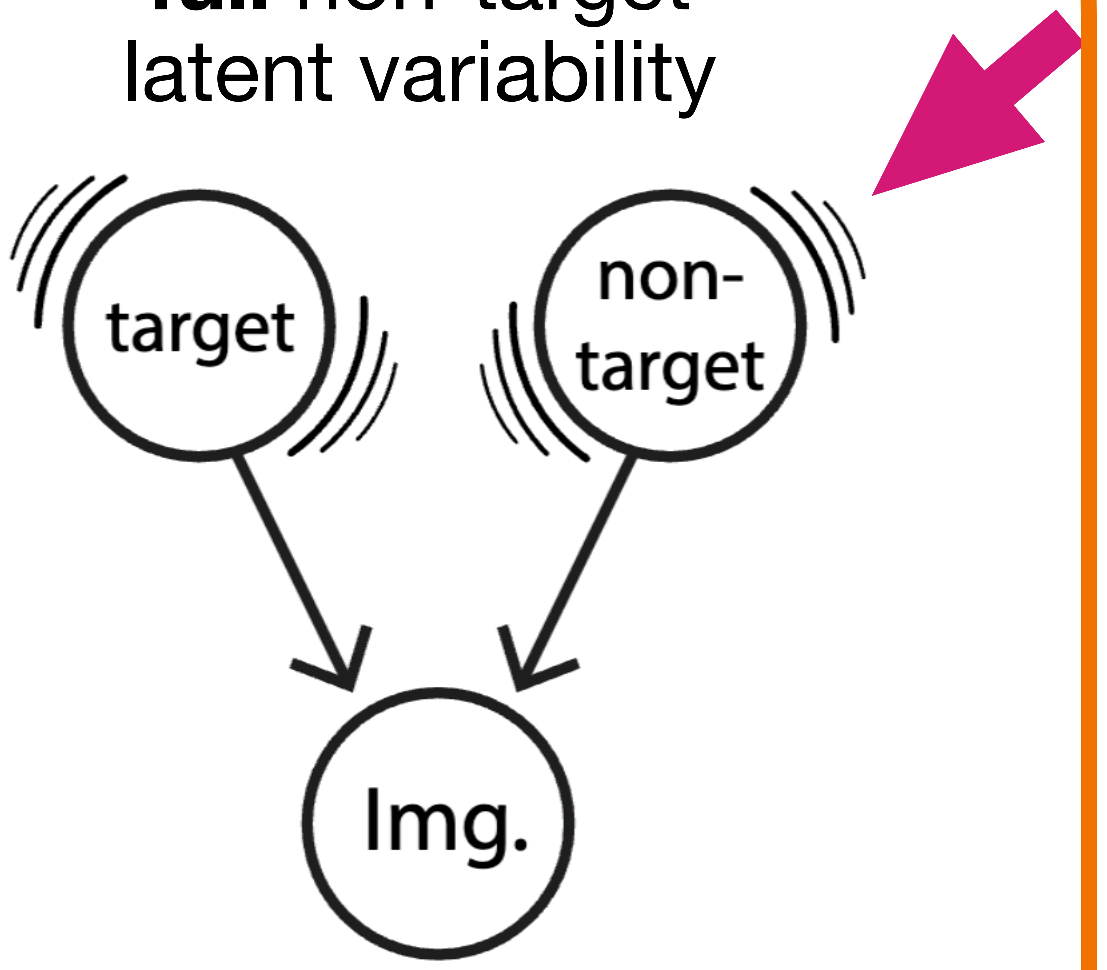
Non-target latent variability helps learn representations of the joint latents

Model training set:
reduced non-target
latent variability



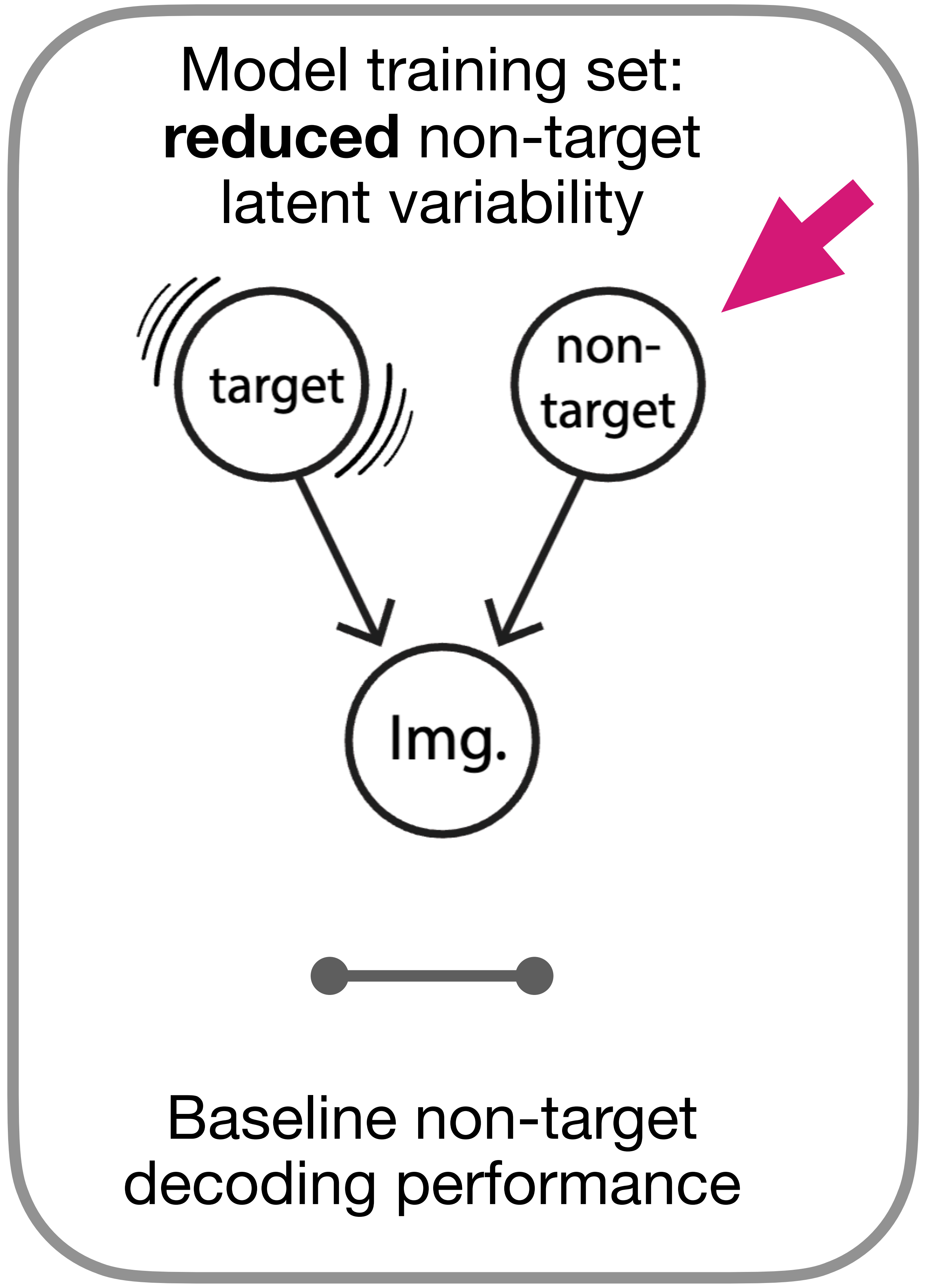
Baseline non-target
decoding performance

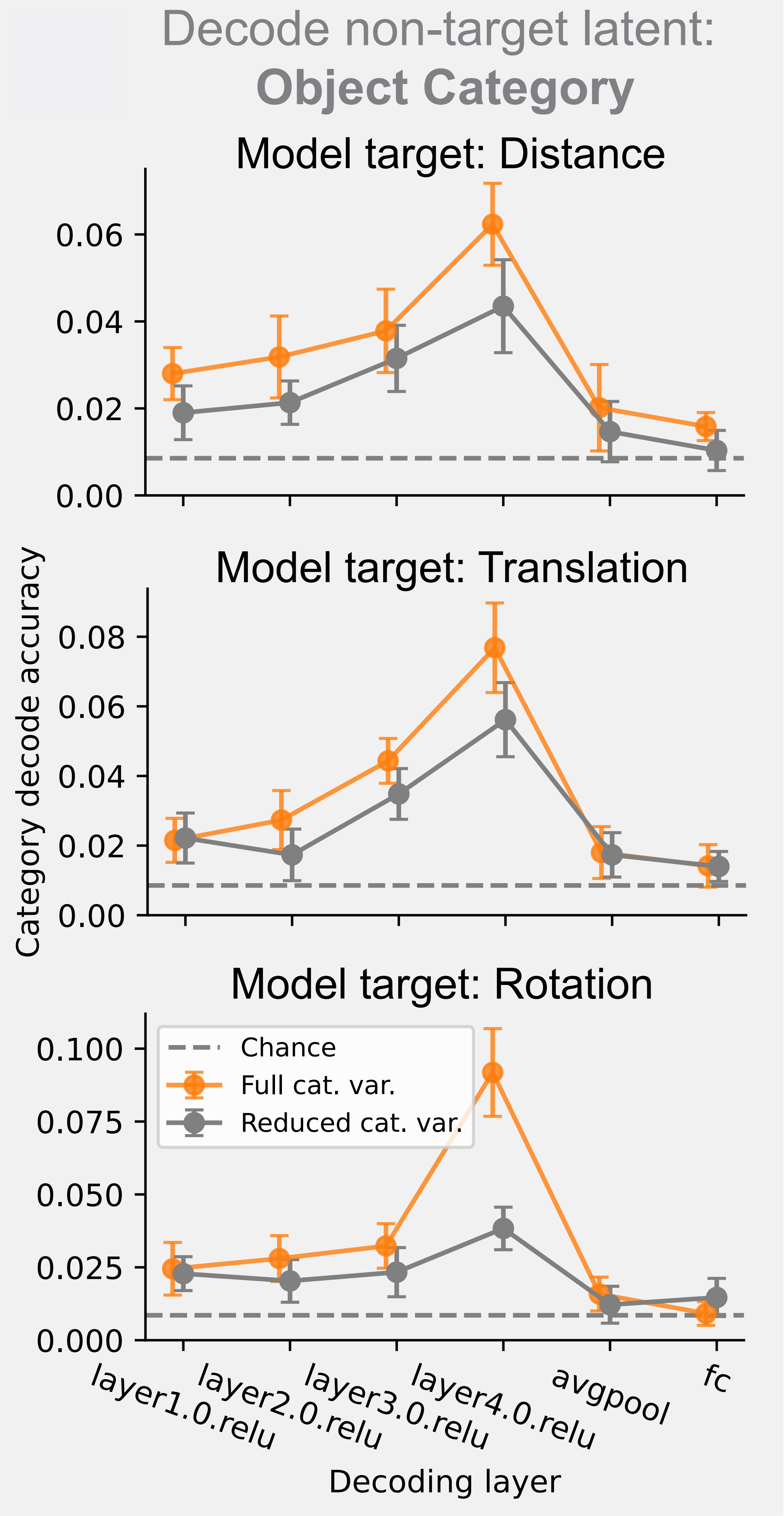
Model training set:
full non-target
latent variability

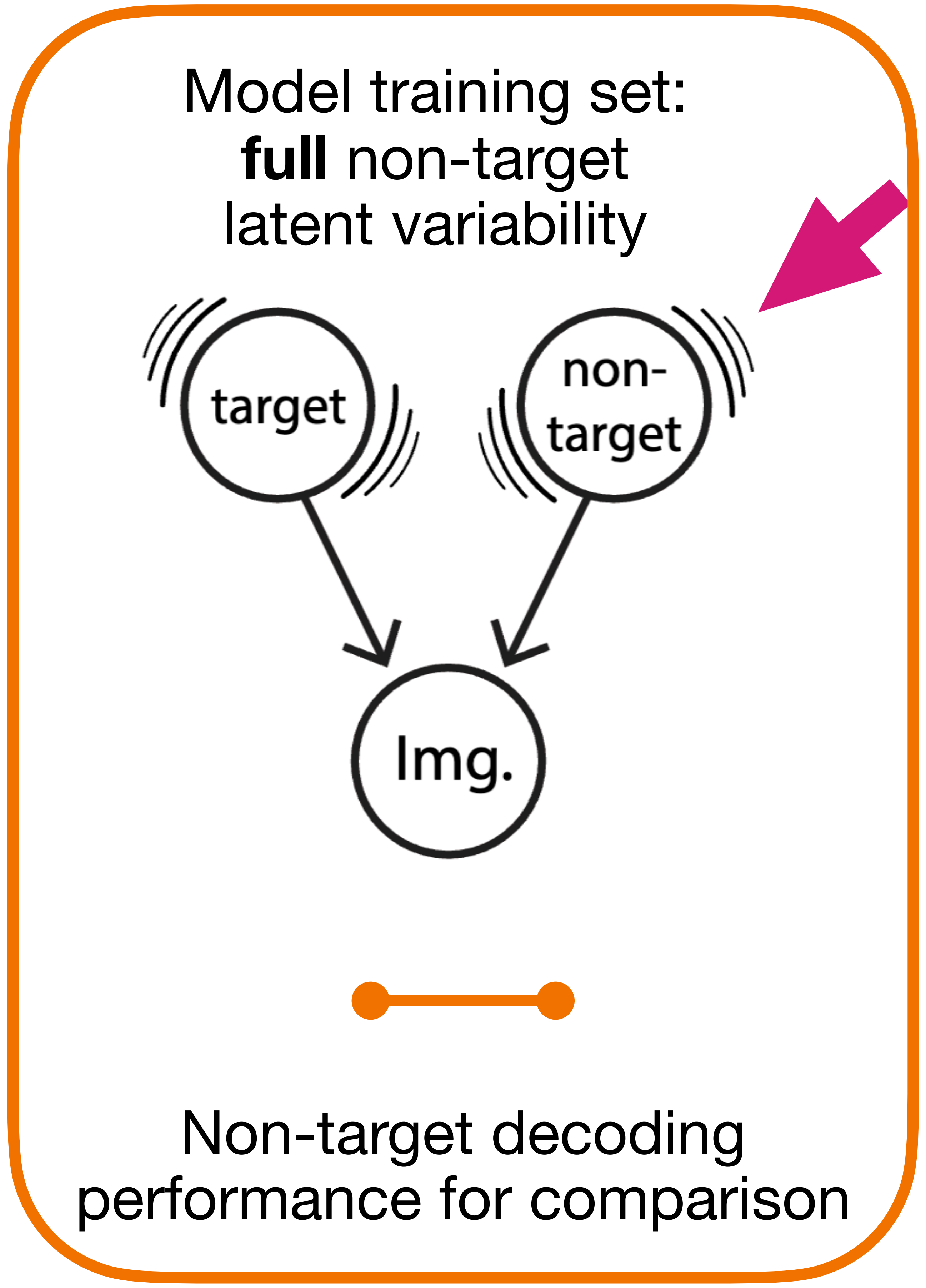


Non-target decoding
performance for comparison

Non-target latent variability helps learn representations of the joint latents

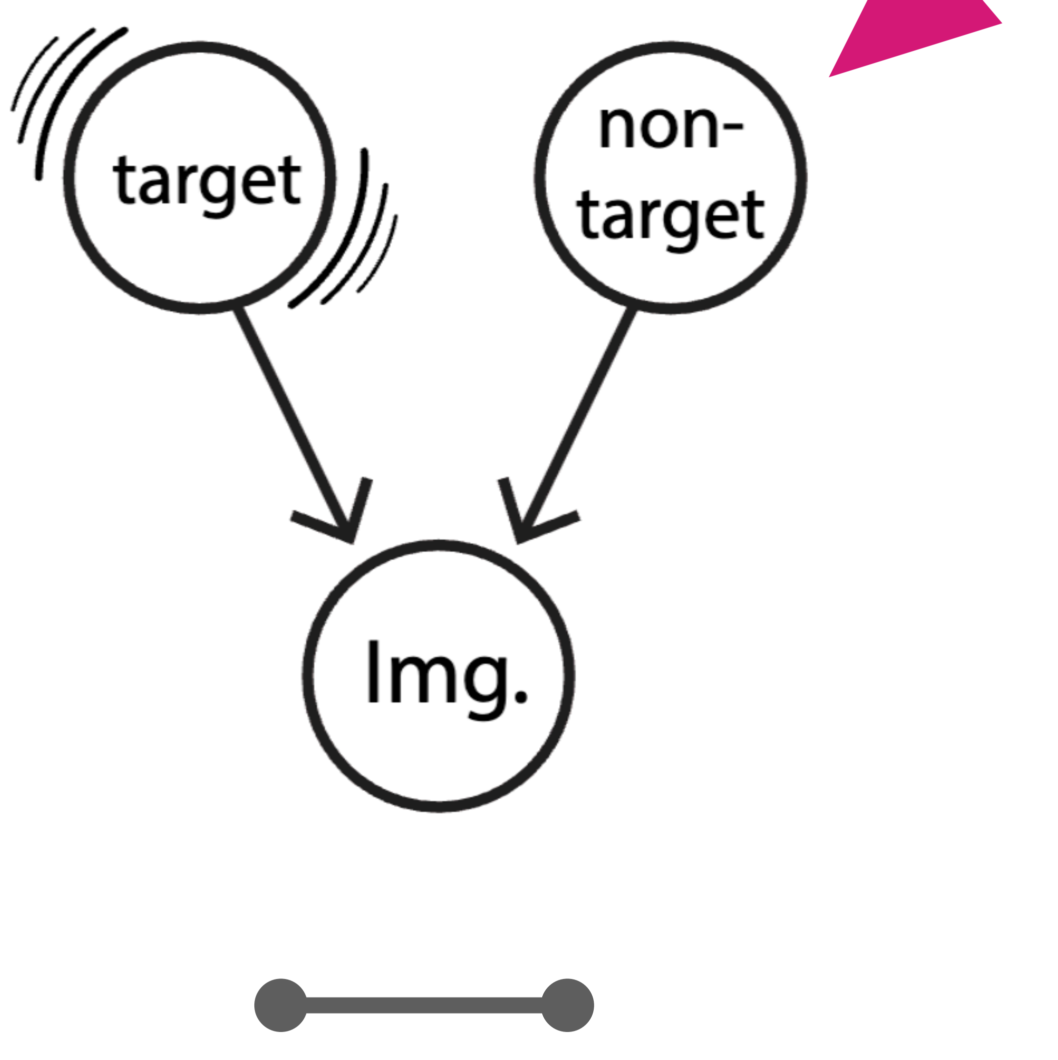




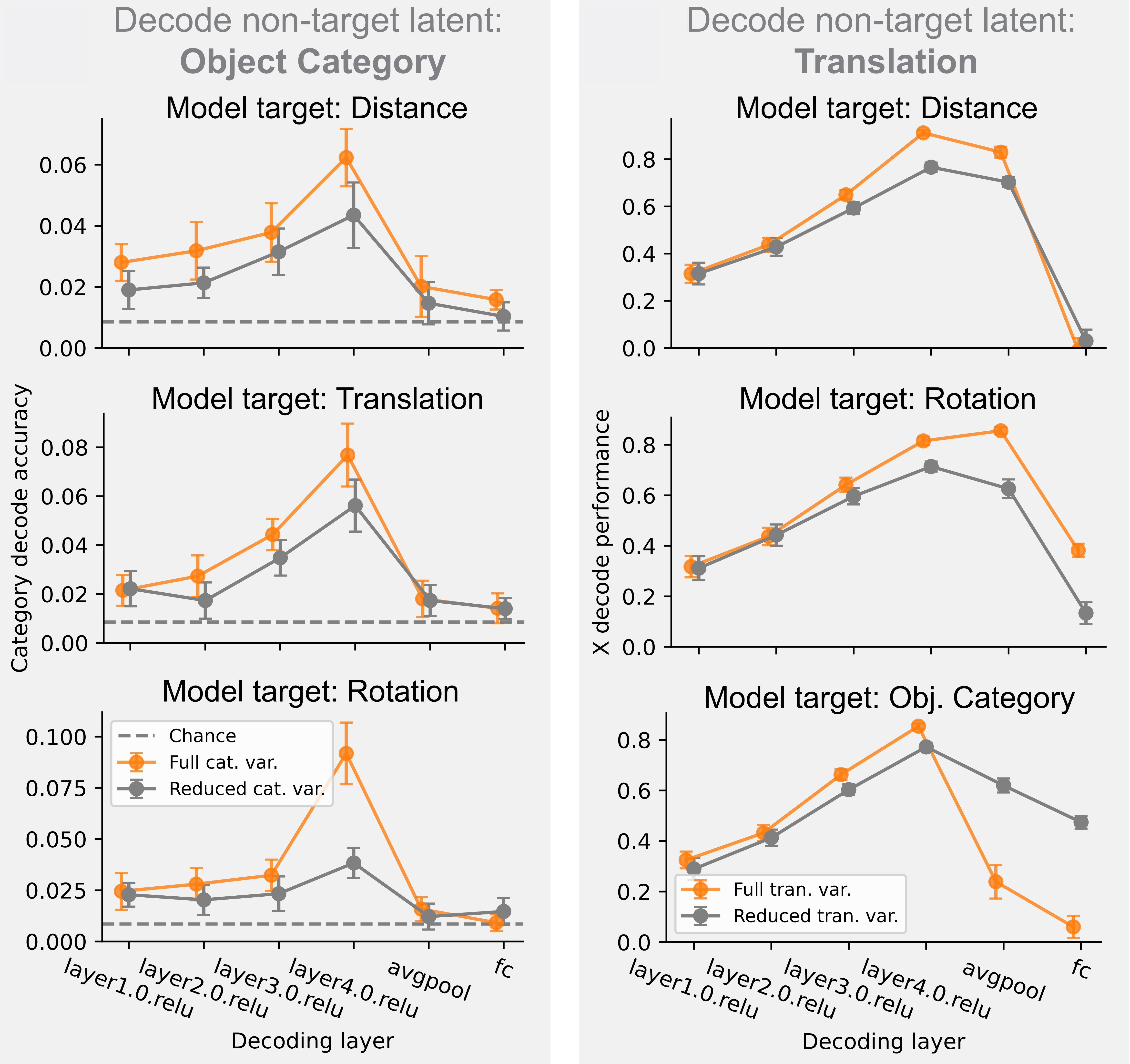


Non-target latent variability helps learn representations of the joint latents

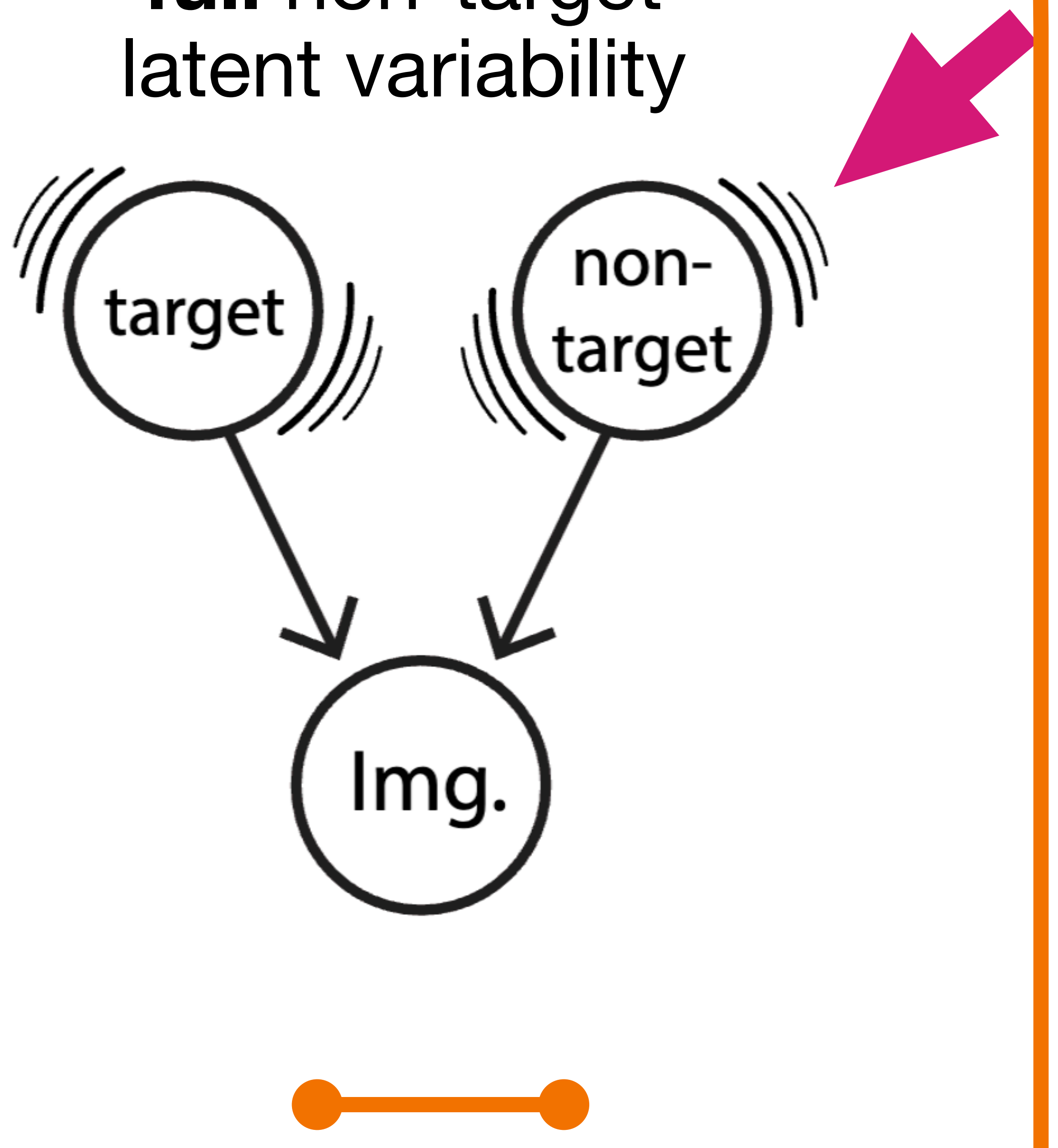
Model training set:
reduced non-target
latent variability



Baseline non-target
decoding performance

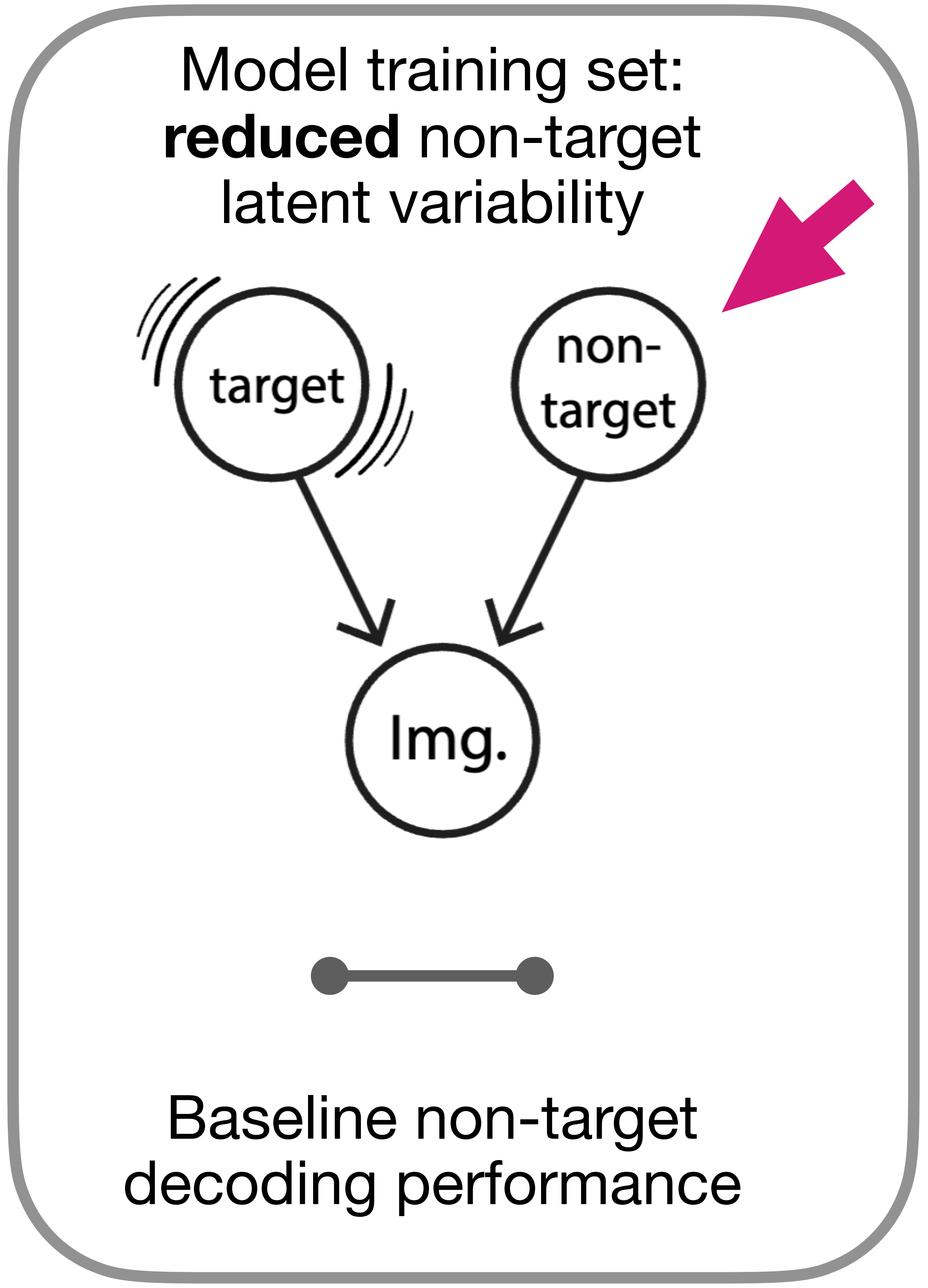


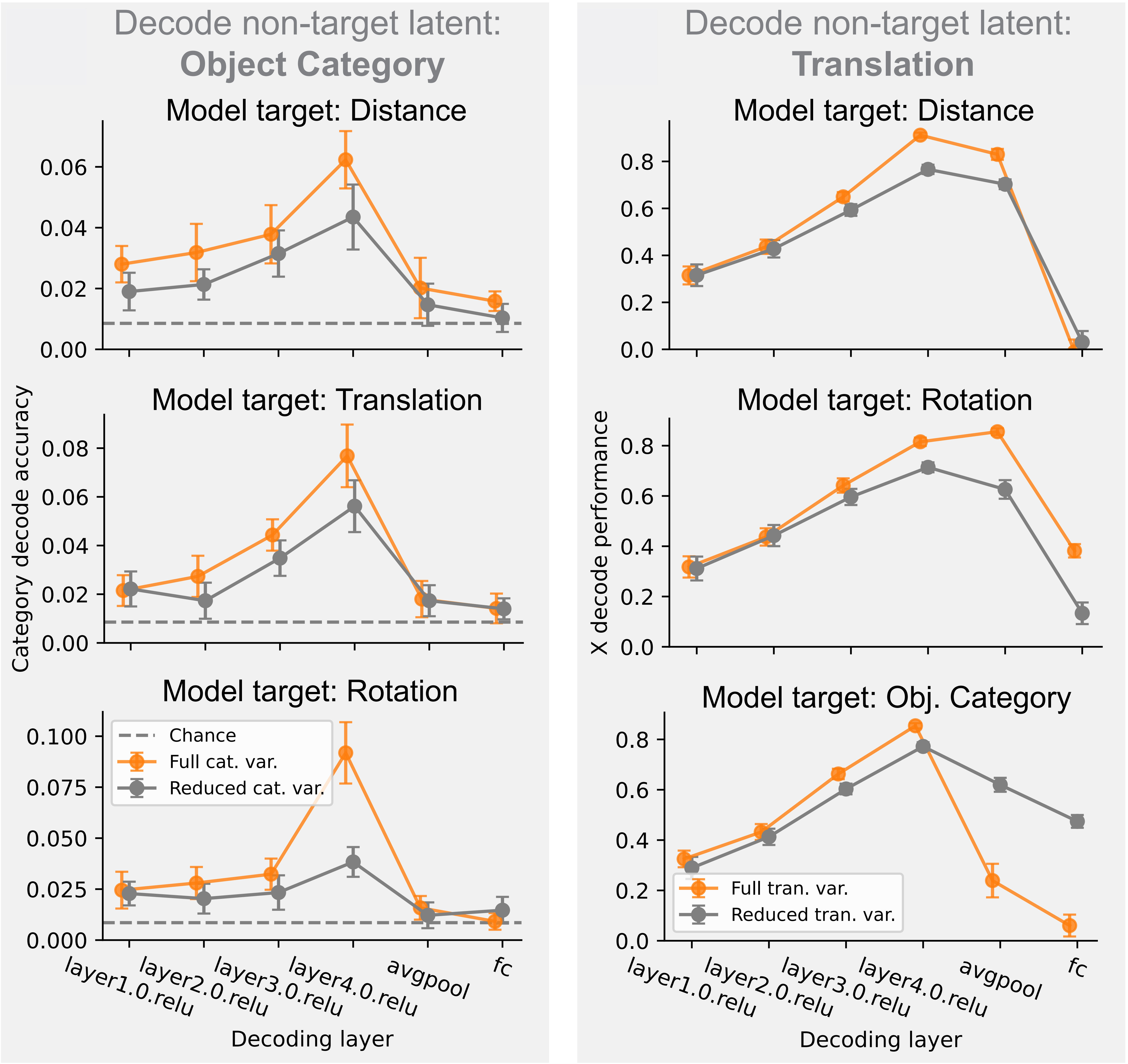
Model training set:
full non-target
latent variability

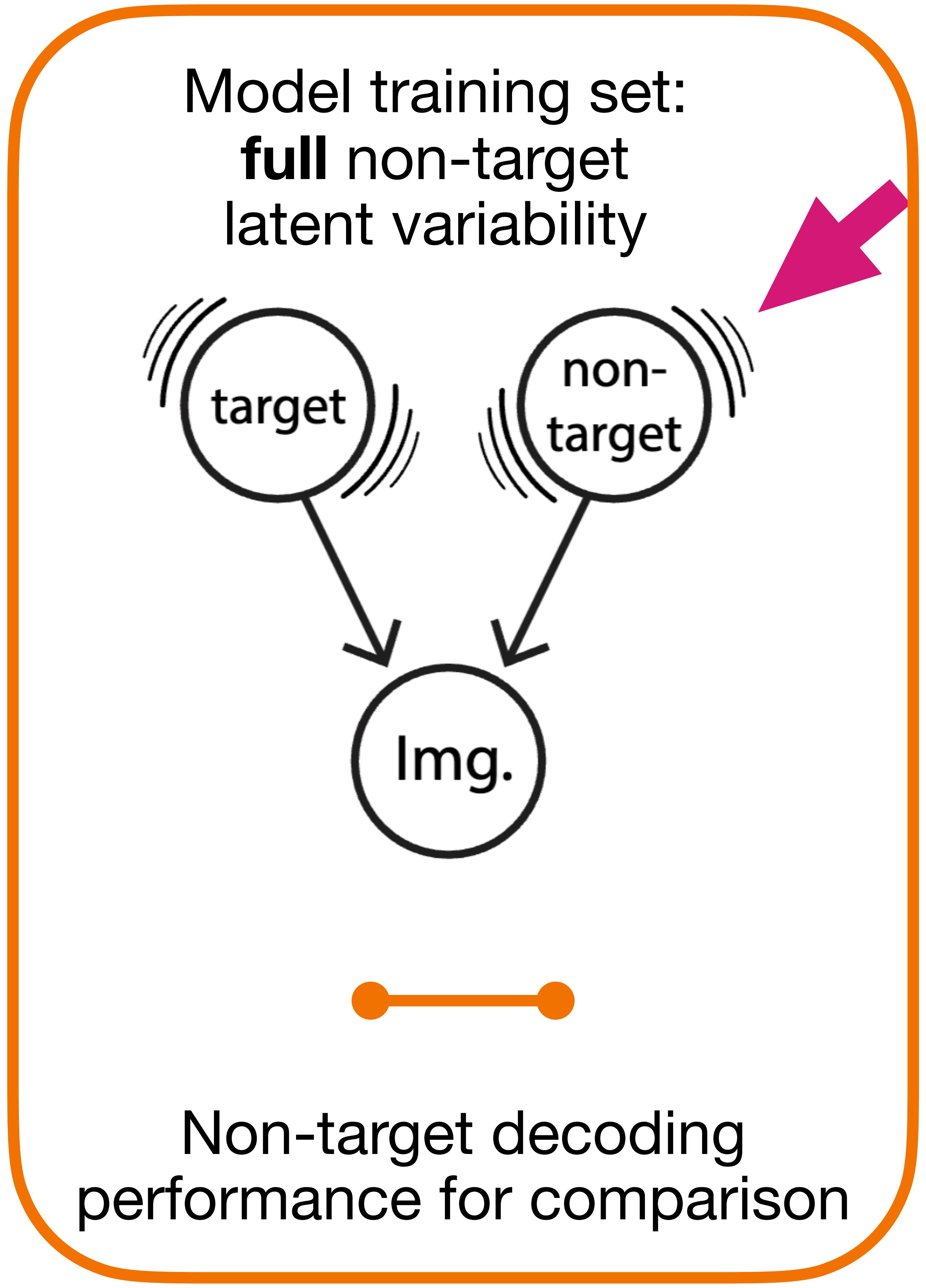


Non-target decoding
performance for comparison

Non-target latent variability helps learn representations of the joint latents

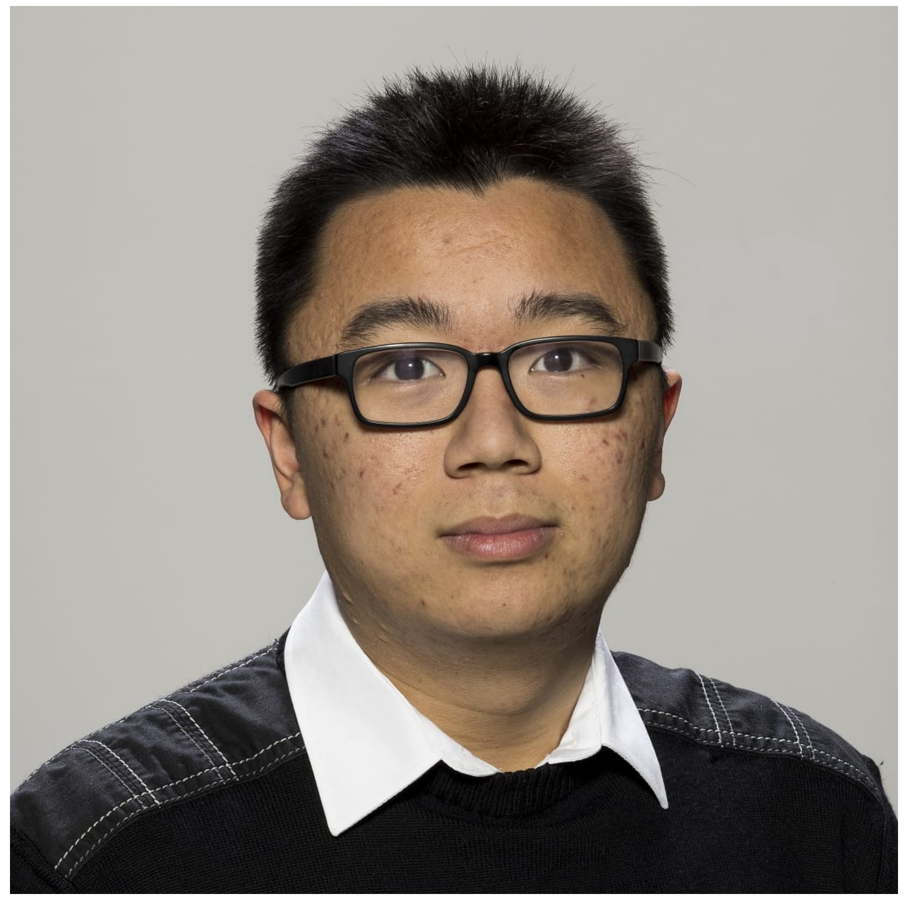






- Added variability of non-target latent helps models learn a better representation of non-target latents.

Acknowledgement


Weichen Huang


Esther Alter


Jeremy Schwartz


Josh Tenenbaum


James DiCarlo