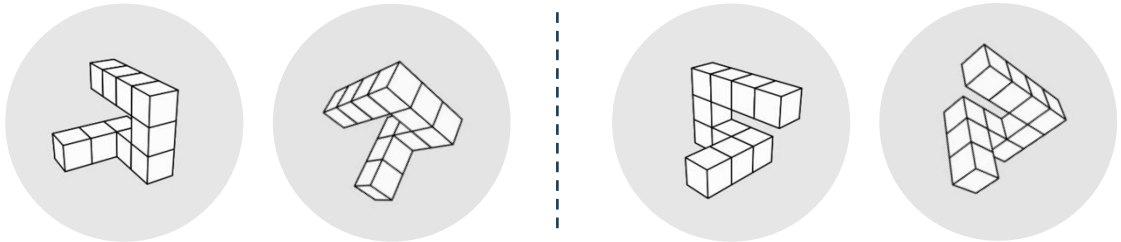
A Deep Learning Model of Mental Rotation Informed by Interactive VR Experiments

Raymond Khazoum, Daniela Fernandes, Aleksandr Krylov, Qin Li, Stéphane Deny

ICML
International Conference
On Machine Learning

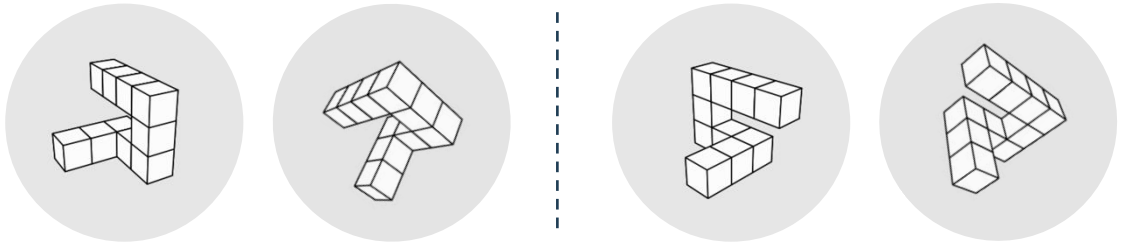
Mental Rotation

A Similarity Assessment Task (same/mirror ?)



Mental Rotation

A Similarity Assessment Task (same/mirror ?)



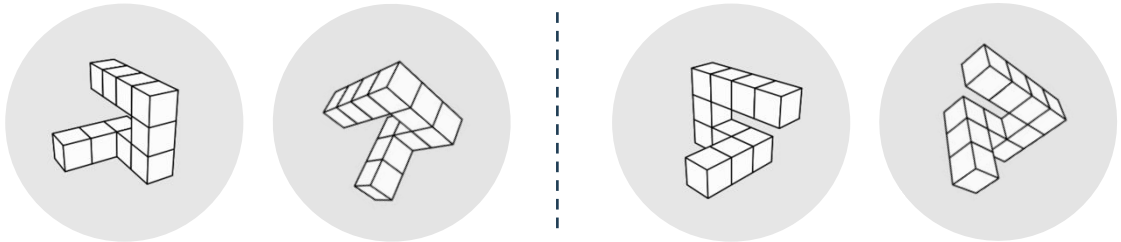
same

mirror

Mental Rotation

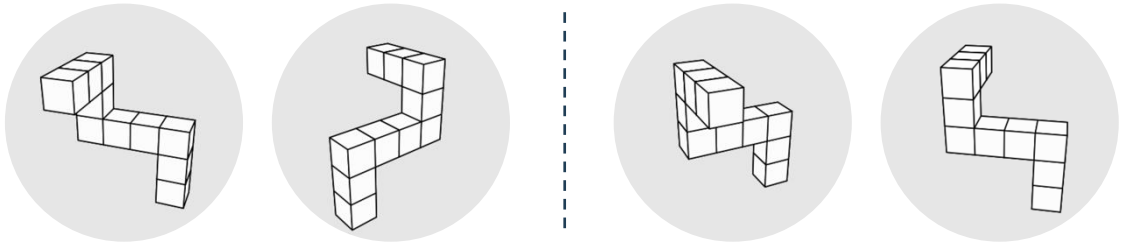
A Similarity Assessment Task (same/mirror ?)

In-plane rotation



same mirror

In-depth rotation

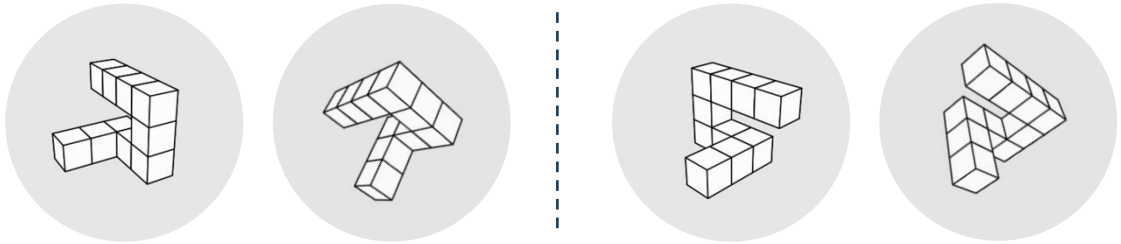


same mirror

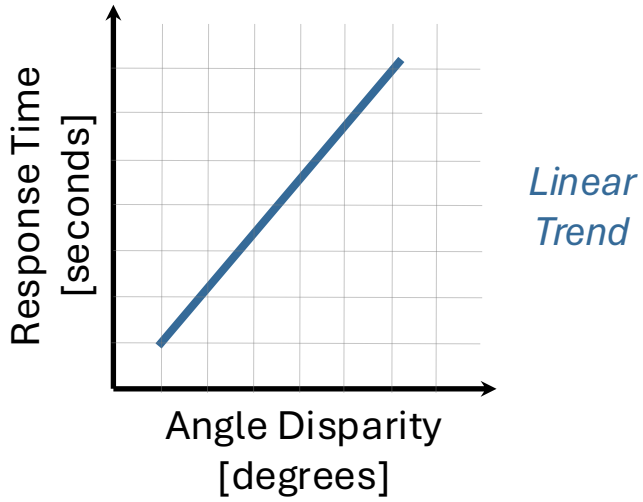
Mental Rotation

The Hallmark Signature [Shepard & Metzler, 1971]

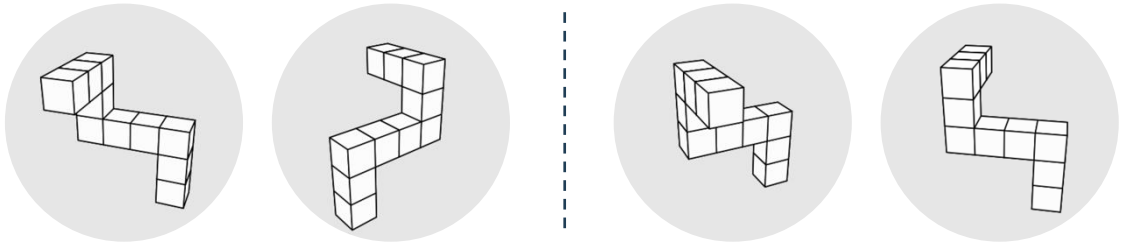
In-plane rotation



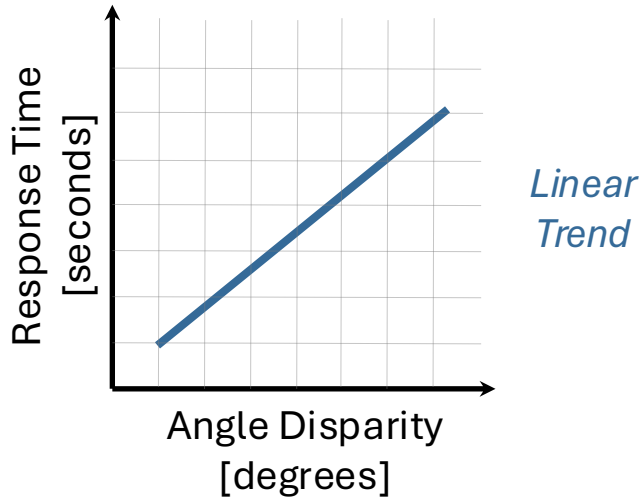
same mirror



In-depth rotation



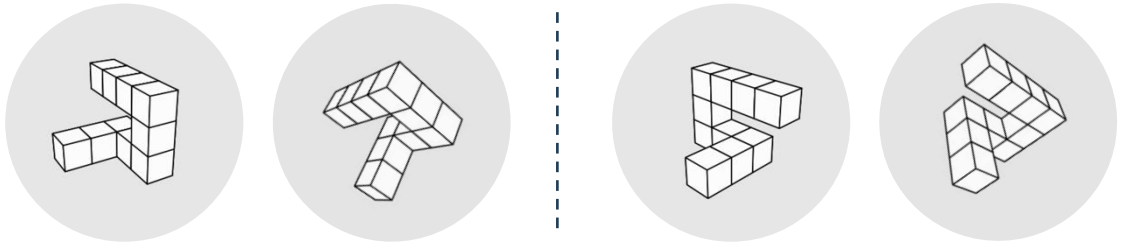
same mirror



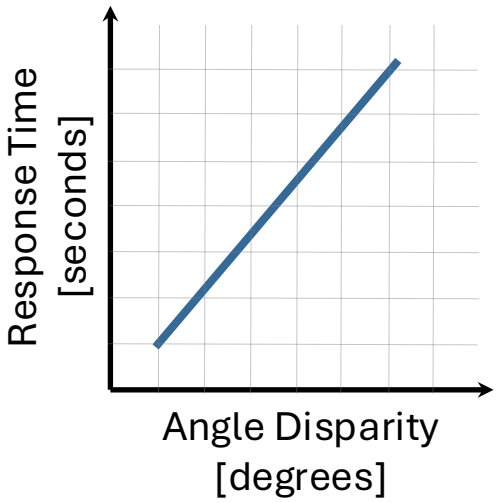
Mental Rotation

The Hallmark Signature [Shepard & Metzler, 1971]

In-plane rotation

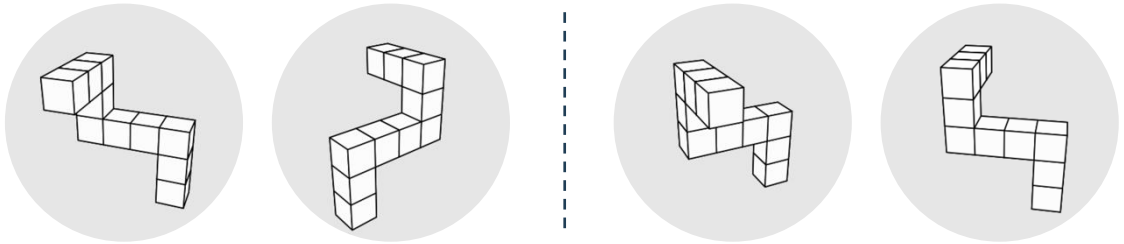


same mirror

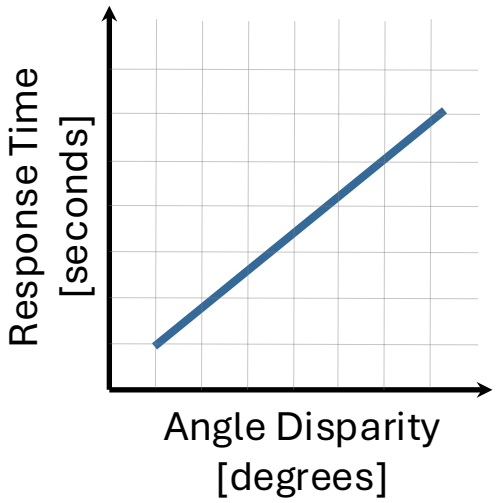


Linear Trend

In-depth rotation



same mirror



Linear Trend

→ *Humans solve **in-depth** and **in-plane** rotations **with similar ease**, as RT scales linearly with angular disparity.*

In This Work

Understanding how humans perform mental rotation, by asking…

Q1: What kind of representation needs to be enforced for human-like spatial reasoning, such as mental rotation, to emerge in deep learning models?

Q2: Do all stages of the cognitive process — rotating, comparing, deciding — operate on the same kind of representation?

In This Work

Understanding how humans perform mental rotation, by asking…

Q1: What kind of representation needs to be enforced for human-like spatial reasoning, such as mental rotation, to emerge in deep learning models?

Q2: Do all stages of the cognitive process — rotating, comparing, deciding — operate on the same kind of representation?

→ *We answer that* **two types of representation** *are needed:*

- **Equivariant**: *captures continuous transformations to support rotation*

- **Symbolic**: *abstracts object pose for similarity judgment*

In This Work

Understanding how humans perform mental rotation, by asking…

Q1: What kind of representation needs to be enforced for human-like spatial reasoning, such as mental rotation, to emerge in deep learning models?

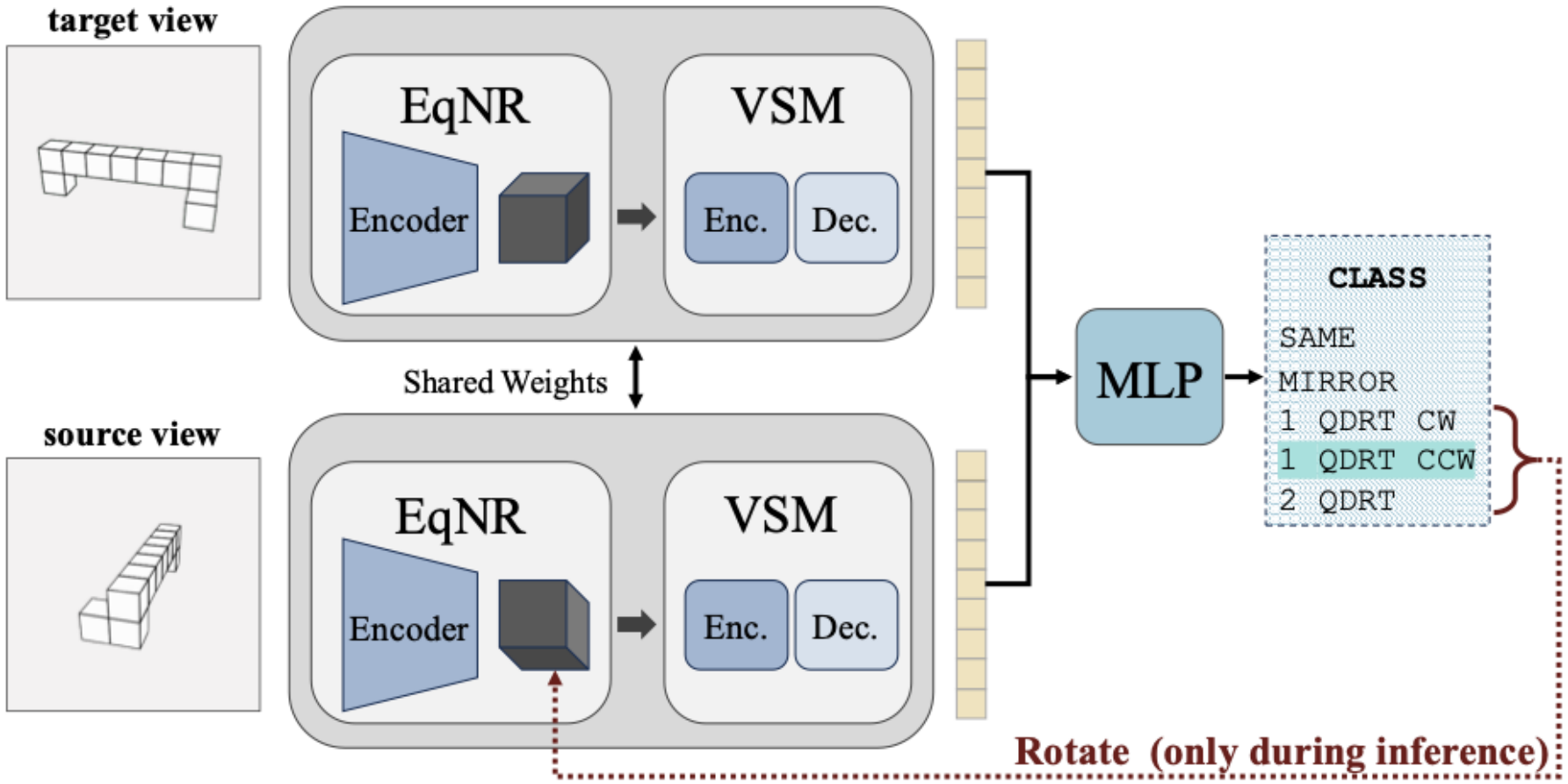
Q2: Do all stages of the cognitive process — rotating, comparing, deciding — operate on the same kind of representation?

→ *We answer that* **two types of representation** *are needed:*

- **Equivariant**: *captures continuous transformations to support rotation*
  → *motivated by the hallmark signature (Linear Trend)*
- **Symbolic**: *abstracts object pose for similarity judgment*
  → *motivated by our interactive VR findings (the Quadrant Hypothesis)*

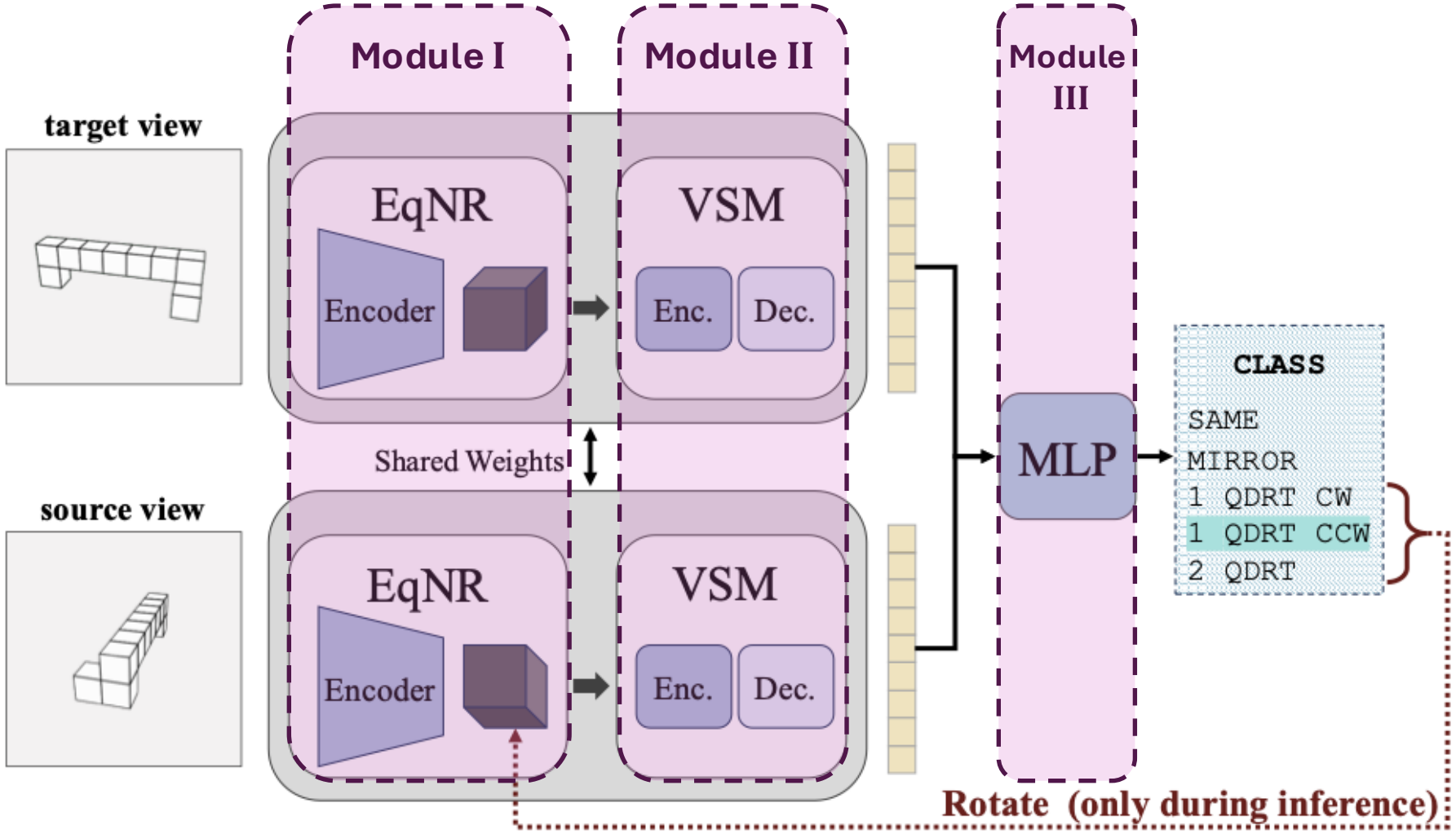
In This Work

We propose a mechanistic model of human mental rotation



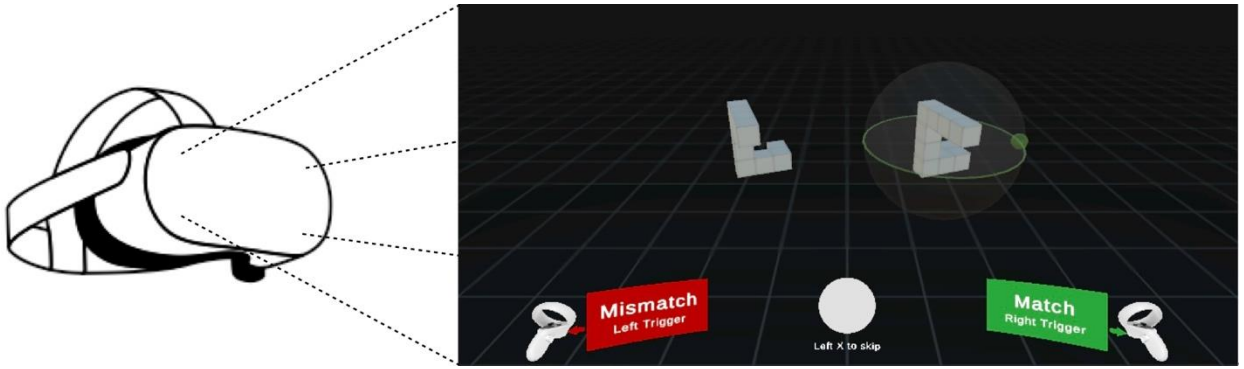
In This Work

We propose a mechanistic model of human mental rotation



In This Work

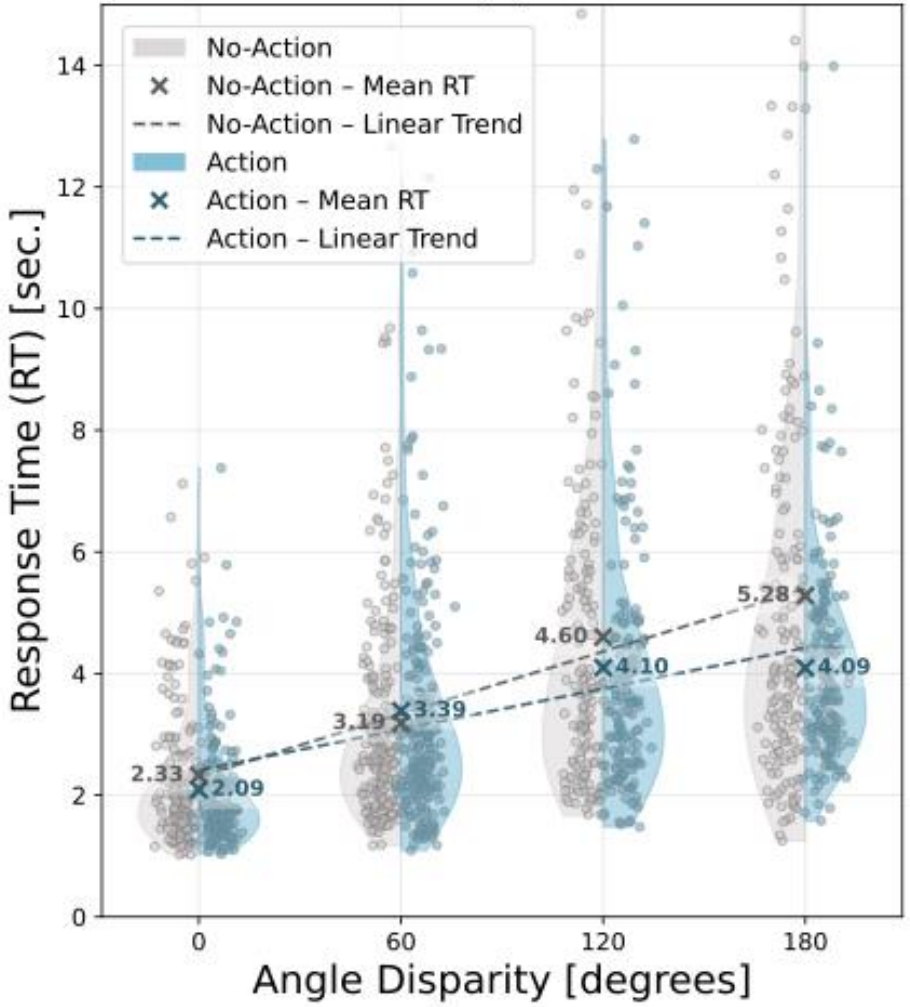
Model design guided by the literature and our VR experiments



➔ *VR app with two operating modes:*

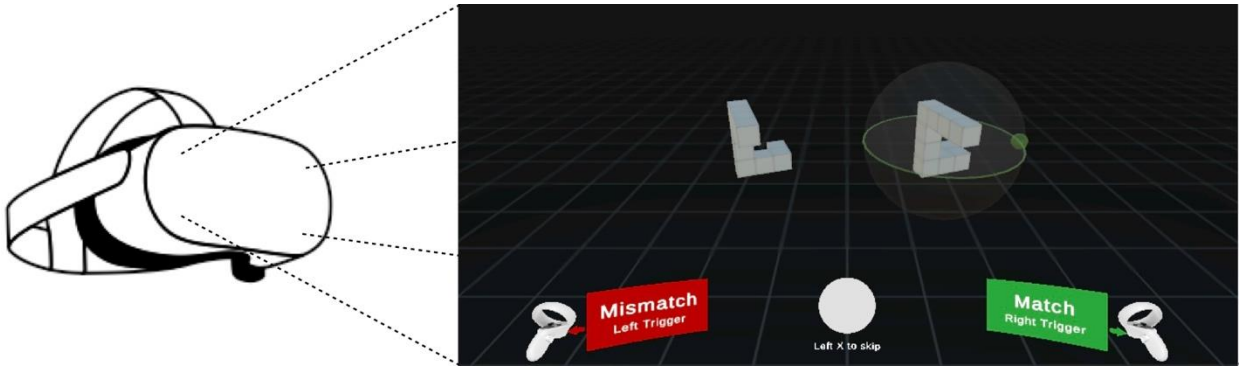
- **No-Action Setup**: *classical Shepard-Metzler mental rotation experiment in 3D, object pairs are presented rotated in depth.*

- **Action Setup**: *same task with the addition that subjects can optionally rotate one of the objects (in depth) using the controller's thumbstick.*



In This Work

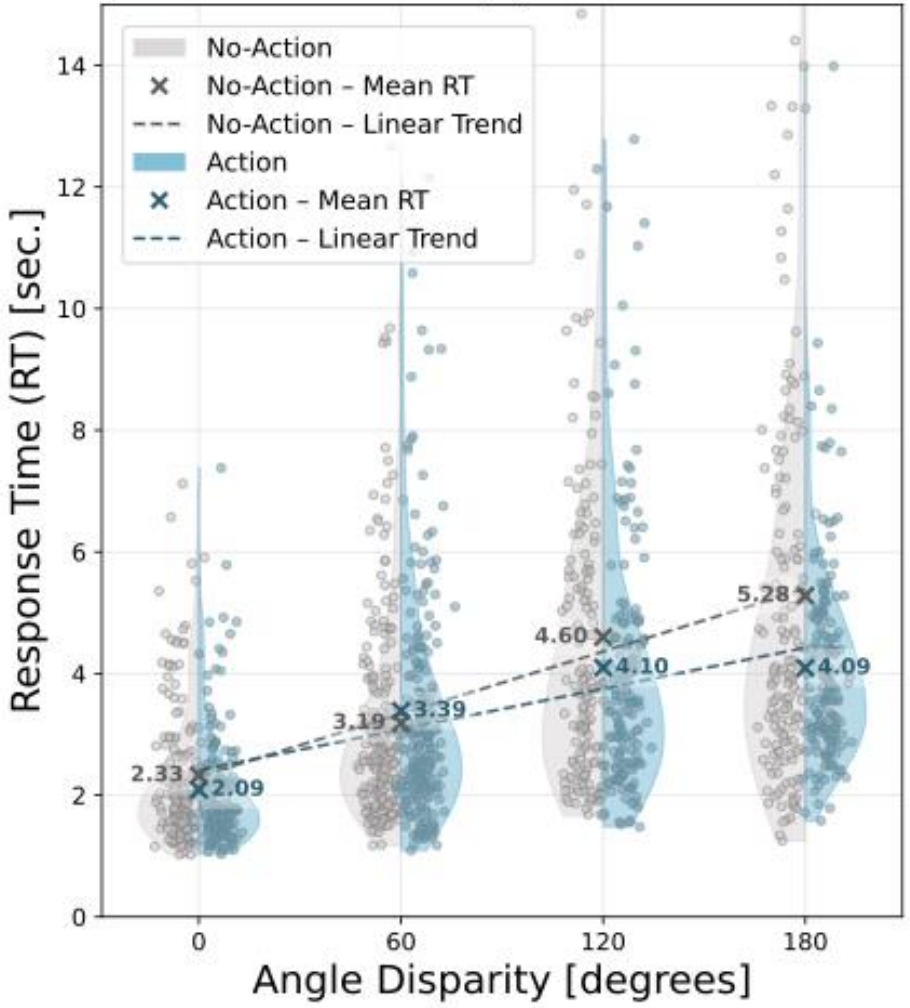
Model design guided by the literature and our VR experiments



➔ *VR app with two operating modes:*

- **No-Action Setup**: *classical Shepard-Metzler mental rotation experiment in 3D, object pairs are presented rotated in depth.*

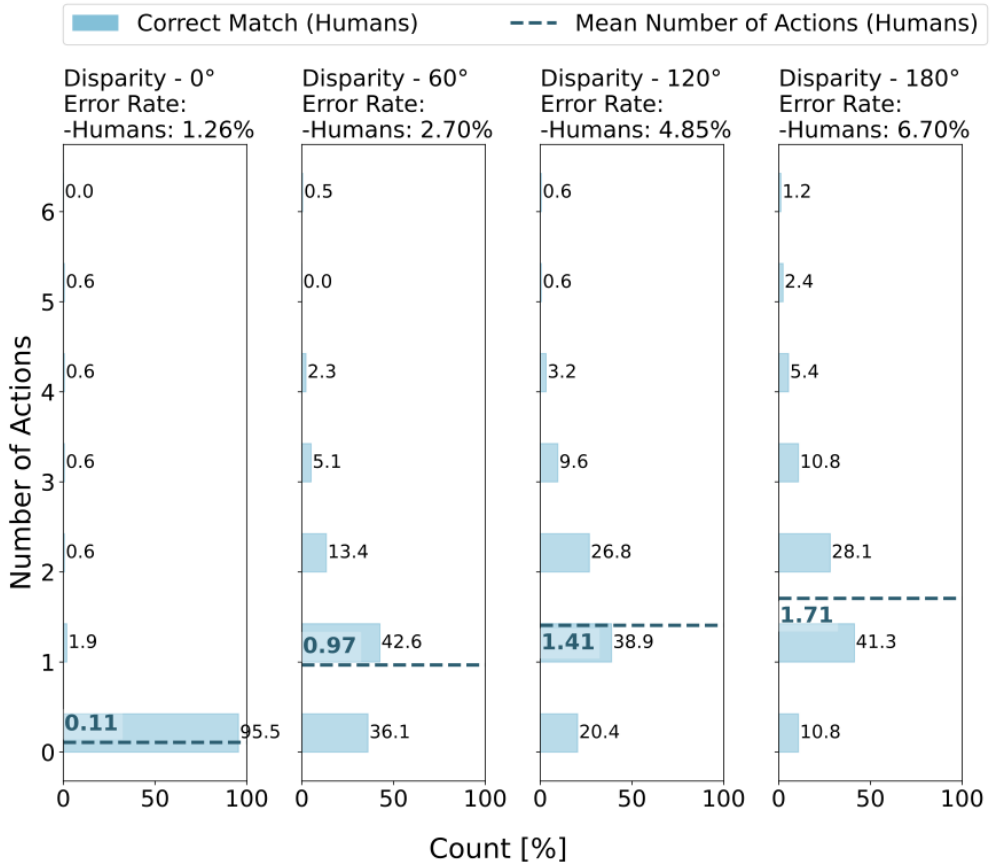
- **Action Setup**: *same task with the addition that subjects can optionally rotate one of the objects (in depth) using the controller's thumbstick.*
  ➔ interactive setup motivated by the need for behavioral signals ***beyond*** *response time.*



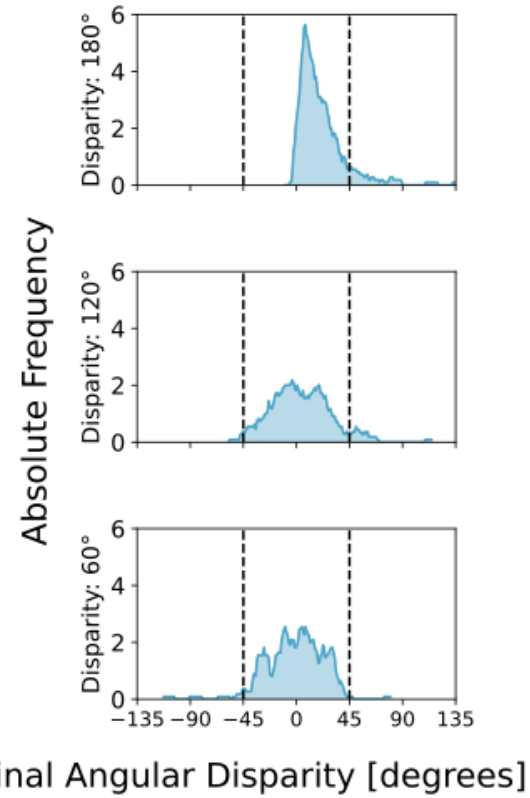
VR Experiments

Findings in the interactive **Action Setup**

→ ***Small number of actions*** *taken by subjects (1.05 on average) before making a decision*

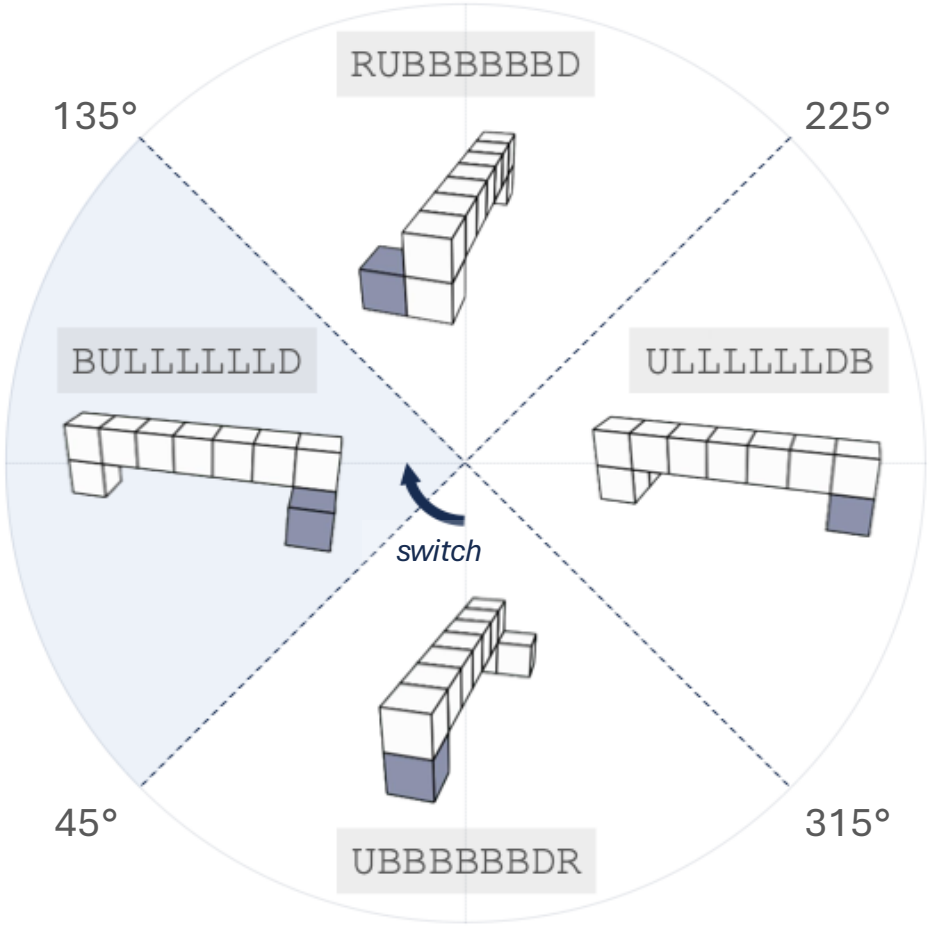


→ *Subjects' tendency to* ***place objects within an angle of [-45°, +45°]*** *relative to each other before making a decision*



The Quadrant Hypothesis

An explanation for interactive VR findings



➔ *Objects are placed into a visual "***quadrant***," abstracting each object's pose to quadrant membership.*

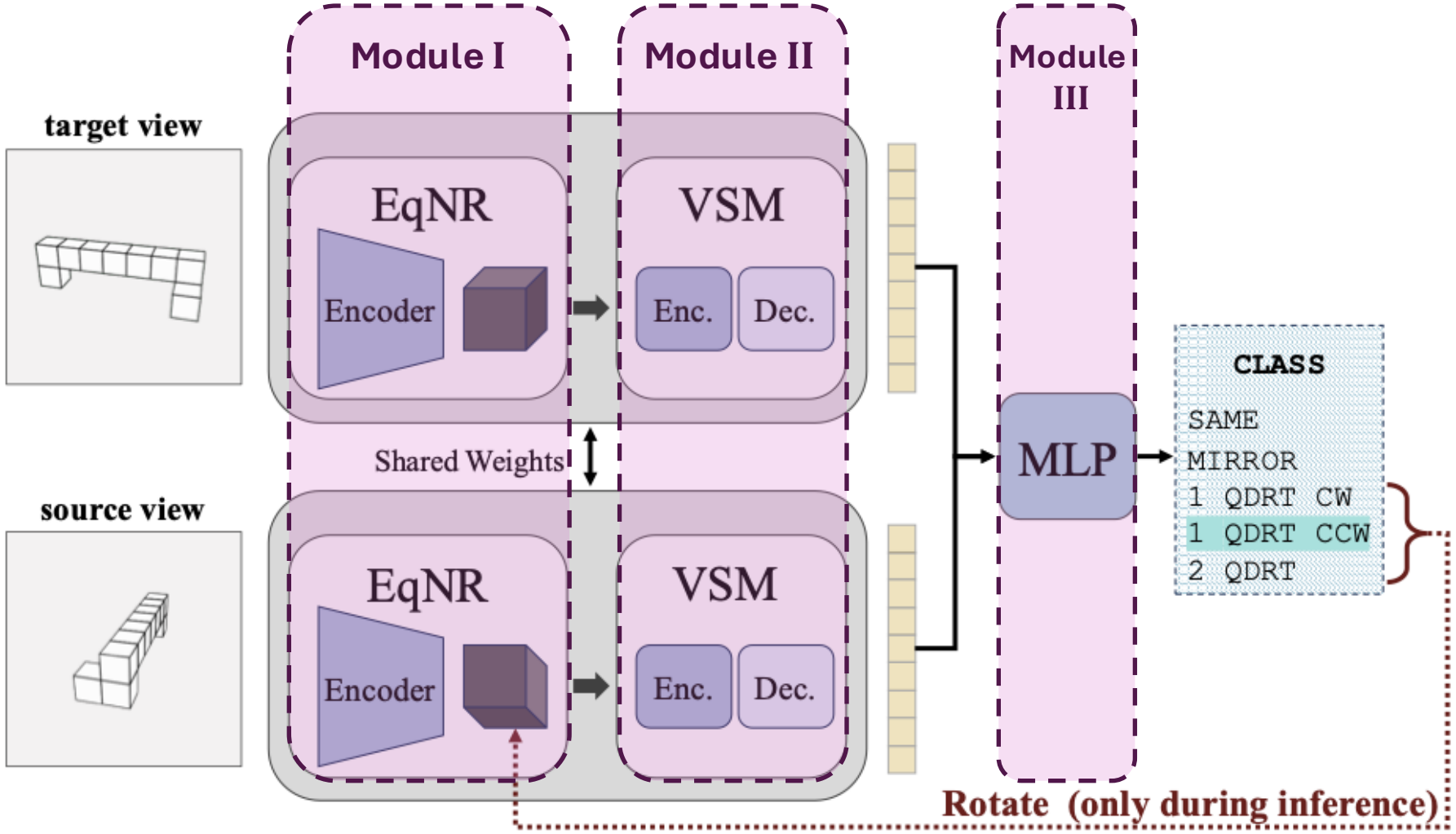
➔ *Each object has a unique code per quadrant; mental rotation* **reduces** *to switching quadrants until alignment.*

➔ **The symbolic code** = *the sequence of transitions between composite cubes (from the nearest cube):*

- `U`: up
- `D`: down
- `B`: back
- `F`: forward
- `L`: left
- `R`: right

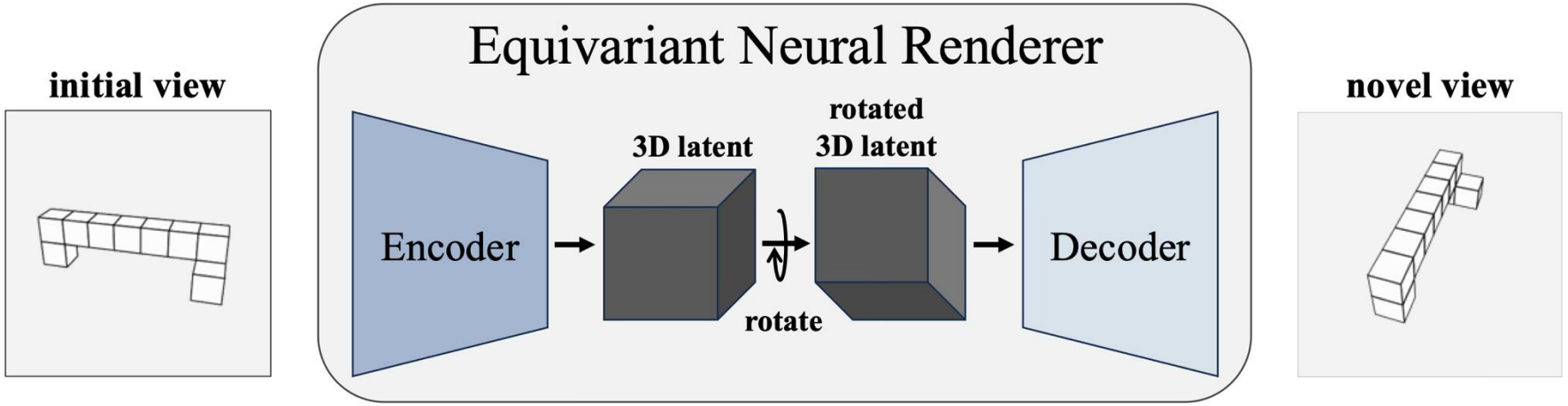
The Model

Composed of three modules



The Model

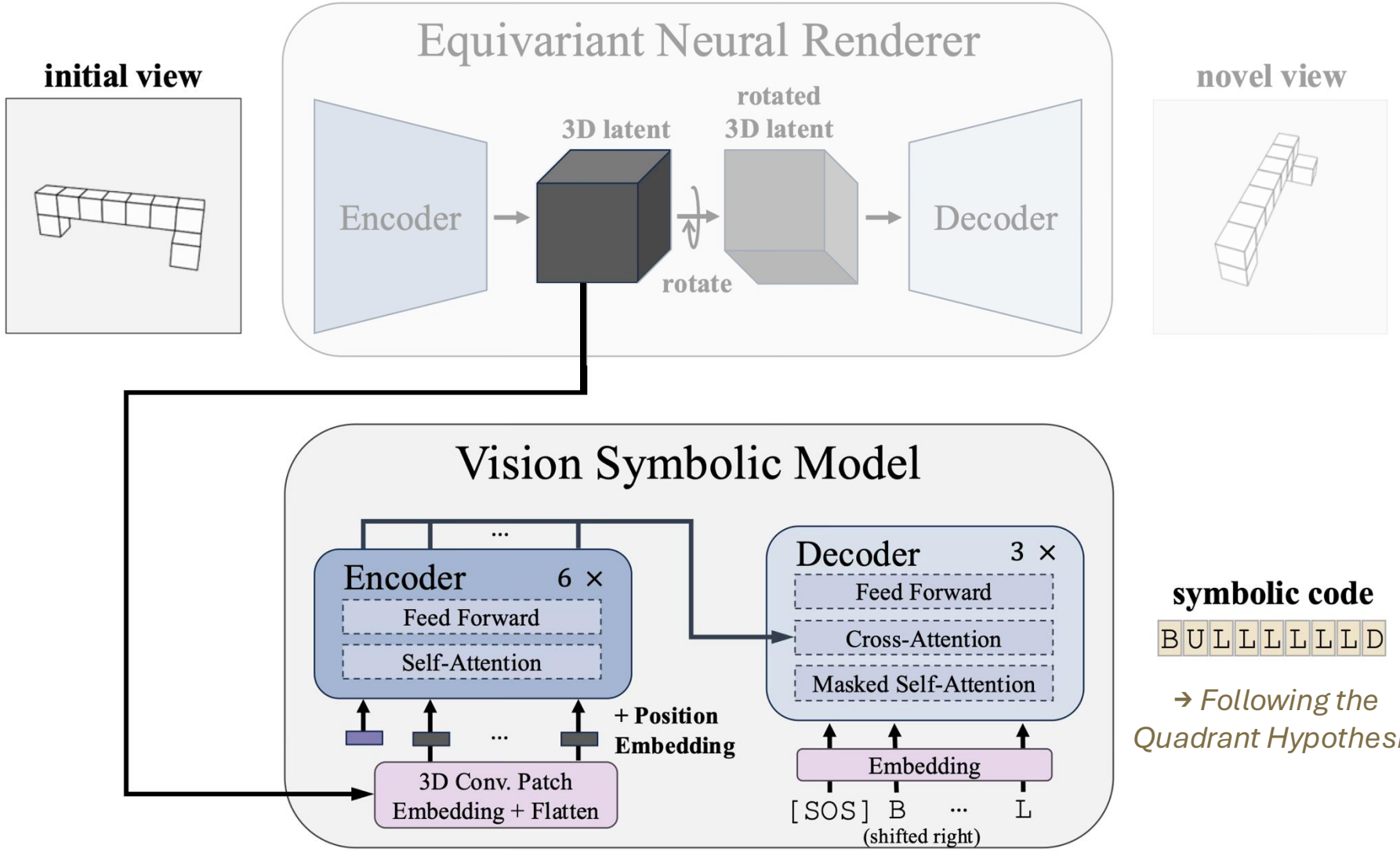
Module I for Spatial Representation



Architecture from [Dupont et al., ICML 2020]

The Model

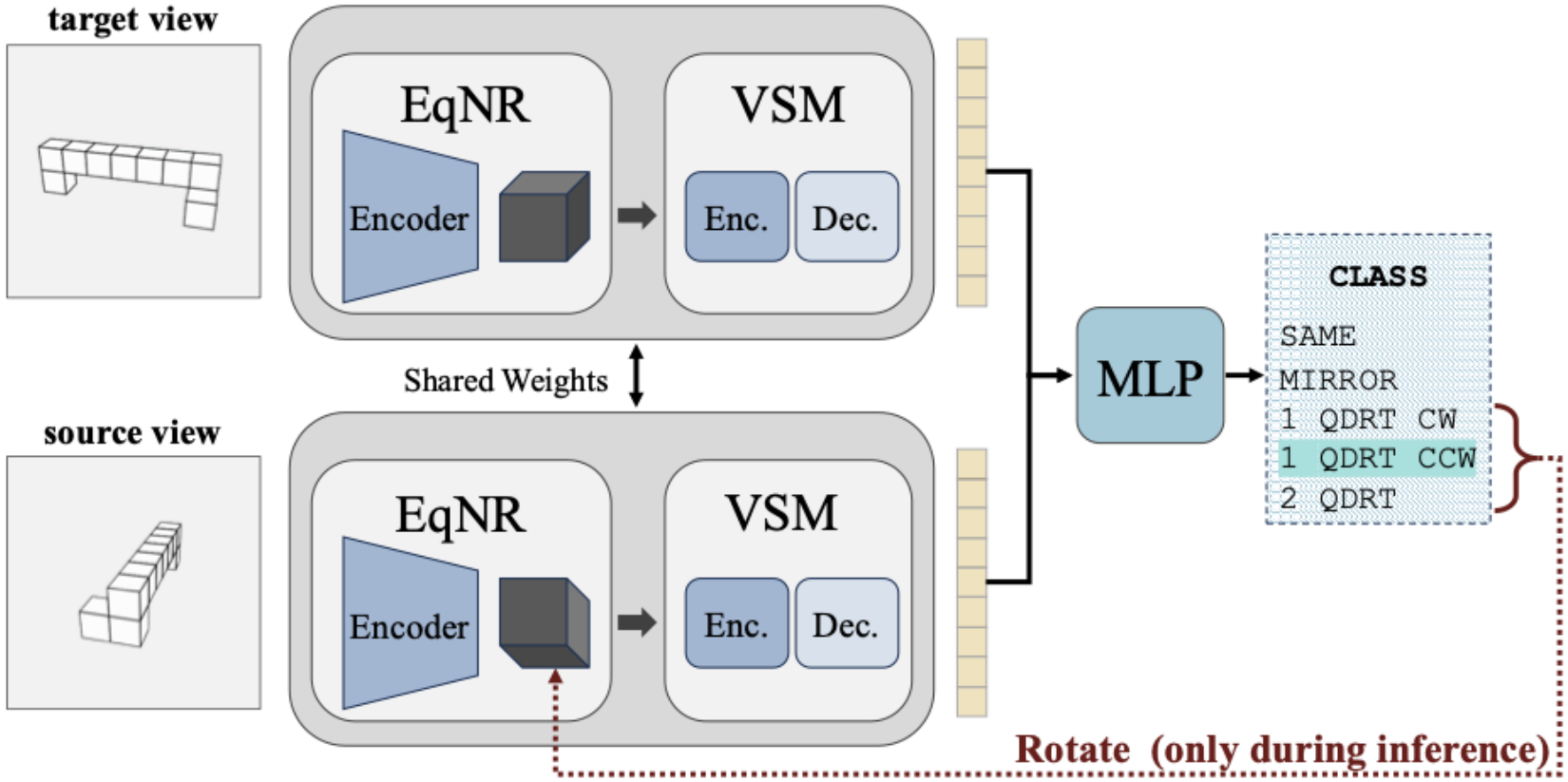
Module II for Symbolic Representation



→ *Following the Quadrant Hypothesis*

The Model

Module III for Similarity Decision and Action Taking



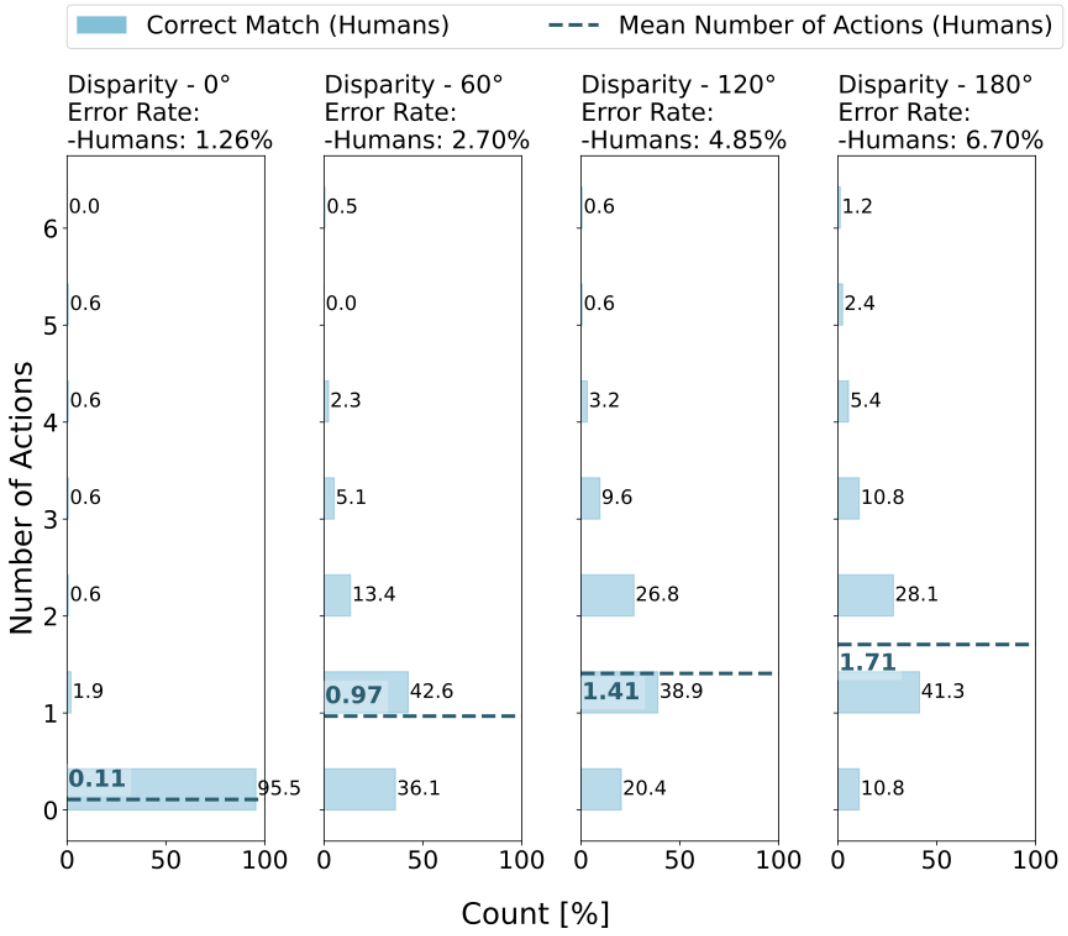
→ *Trained as a supervised 5-class problem — predicting a* **similarity decision** (`same`/`mirror`) *or a* **rotation action** (90°: `1 QDRT CW`, −90°: `1 QDRT CCW`, 180°: `2 QDRT`)

→ *Sequential rotation action-taking is emergent at* **inference** *— iterate until a similarity decision;* **fail** *if no decision after 6 actions*

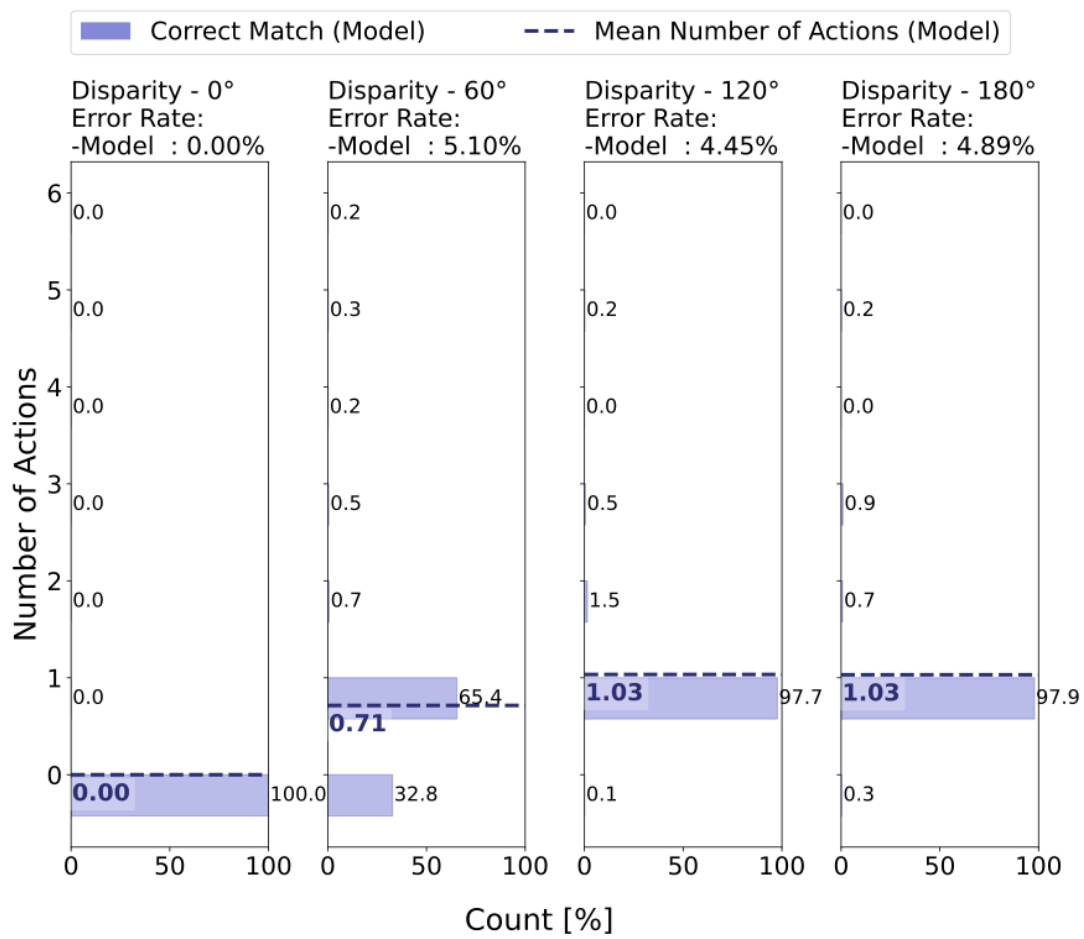
Model Validation

Performance & Comparison with Human Behavior on the Mental Rotation Task

Accuracy: 95.33%

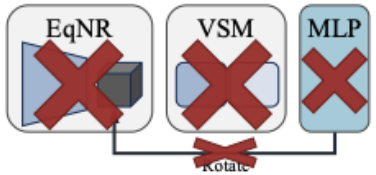


Accuracy: 96.13%



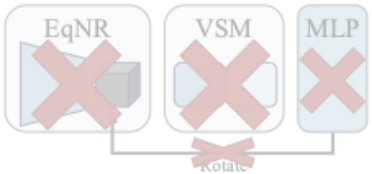
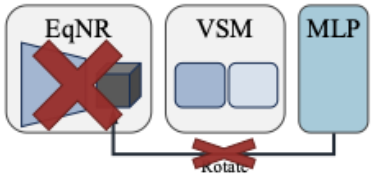
Model Validation

Ablation Study

Model	*Schematic*	*Accuracy*	*Number of Actions*
Siamese Encoders (Baseline)		~ 50% (chance level) on test objects	N/A (no actions taken)

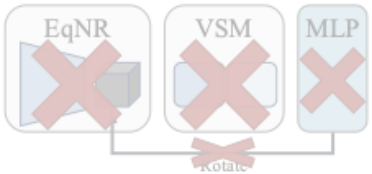
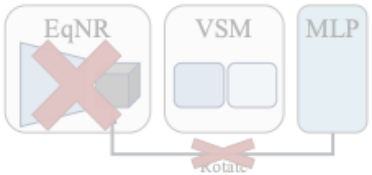
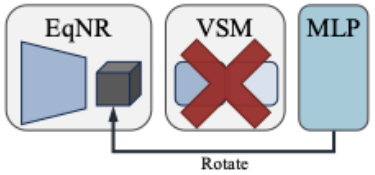
Model Validation

Ablation Study

Model	*Schematic*	*Accuracy*	*Number of Actions*
Siamese Encoders (Baseline)		~ 50% (chance level) on test objects	N/A (no actions taken)
Module I Ablation (w/o Equivariant Rep.)		N/A (the symbolic module fails to encode the object)	N/A (no actions taken)

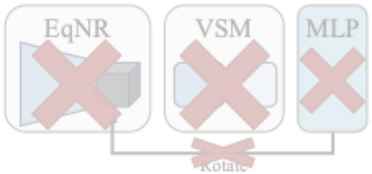
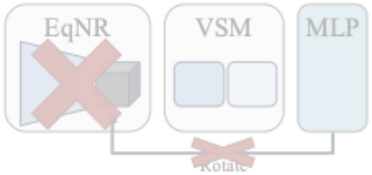
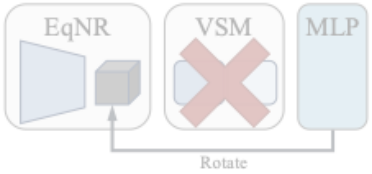
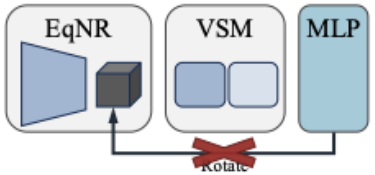
Model Validation

Ablation Study

Model	*Schematic*	*Accuracy*	*Number of Actions*
Siamese Encoders (Baseline)		~ 50% (chance level) on test objects	N/A (no actions taken)
Module I Ablation (w/o Equivariant Rep.)		N/A (the symbolic module fails to encode the object)	N/A (no actions taken)
Module II Ablation (w/o Symbolic Rep.)		~ 90% (If ViT retained), ~ 38% (Otherwise)	Small numbers of actions taken, performances drop for some conditions.

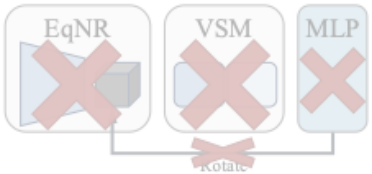
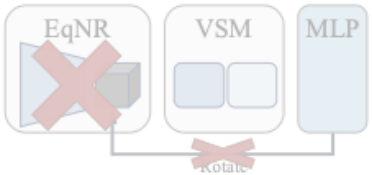
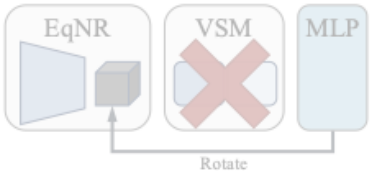
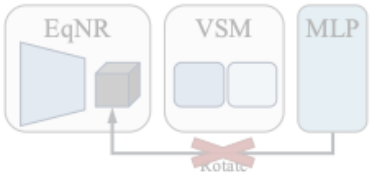
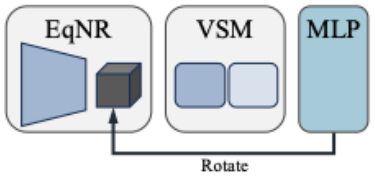
Model Validation

Ablation Study

Model	*Schematic*	*Accuracy*	*Number of Actions*
Siamese Encoders (Baseline)		~ 50% (chance level) on test objects	N/A (no actions taken)
Module I Ablation (w/o Equivariant Rep.)		N/A (the symbolic module fails to encode the object)	N/A (no actions taken)
Module II Ablation (w/o Symbolic Rep.)		~ 90% (If ViT retained), ~ 38% (Otherwise)	Small numbers of actions taken, performances drop for some conditions.
Module III Ablation (w/o Actions)		~ 97% on test objects	N/A (no actions taken)

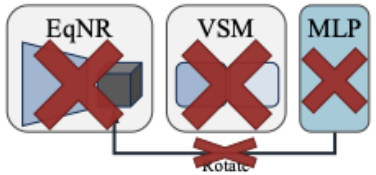
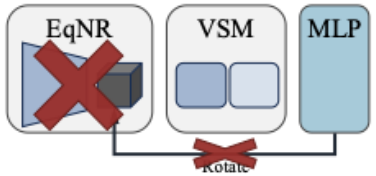
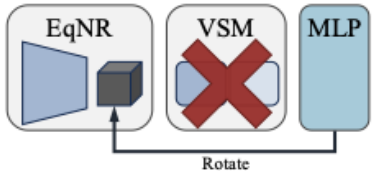
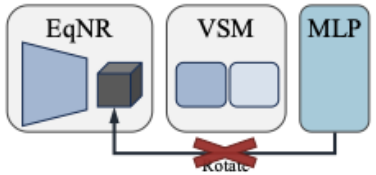
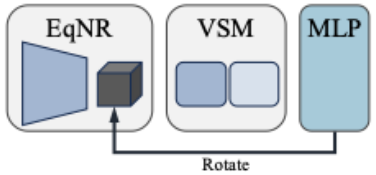
Model Validation

Ablation Study

Model	*Schematic*	*Accuracy*	*Number of Actions*
Siamese Encoders (Baseline)		~ 50% (chance level) on test objects	N/A (no actions taken)
Module I Ablation (w/o Equivariant Rep.)		N/A (the symbolic module fails to encode the object)	N/A (no actions taken)
Module II Ablation (w/o Symbolic Rep.)		~ 90% (If ViT retained), ~ 38% (Otherwise)	Small numbers of actions taken, performances drop for some conditions.
Module III Ablation (w/o Actions)		~ 97% on test objects	N/A (no actions taken)
Ours (Full Model)		~ 96% on test objects	Small numbers of actions taken, compatible with humans.

Model Validation

Ablation Study

Model	*Schematic*	*Accuracy*	*Number of Actions*
Siamese Encoders (Baseline)		~ 50% (chance level) on test objects	N/A (no actions taken)
Module I Ablation (w/o Equivariant Rep.)		N/A (the symbolic module fails to encode the object)	N/A (no actions taken)
Module II Ablation (w/o Symbolic Rep.)		~ 90% (If ViT retained), ~ 38% (Otherwise)	Small numbers of actions taken, performances drop for some conditions.
Module III Ablation (w/o Actions)		~ 97% on test objects	N/A (no actions taken)
Ours (Full Model)		~ 96% on test objects	Small numbers of actions taken, compatible with humans.

Conclusion

Novelty:
We propose a mechanistic model of human mental rotation

- ➔ *reproduces humans' ability to compare pairs of Shepard-Metzler objects*
- ➔ *captures humans' performance and behavioral patterns*

Representations:
Equivariant and symbolic components to model human spatial reasoning

Conclusion

Novelty:
We propose a mechanistic model of human mental rotation

- ➔ *reproduces humans' ability to compare pairs of Shepard-Metzler objects*
- ➔ *captures humans' performance and behavioral patterns*

Representations:
Equivariant and symbolic components to model human spatial reasoning

Limitations:

- *Trained only on Shepard-Metzler objects*
- *Stimuli cover only rotation in depth (Y-axis)*