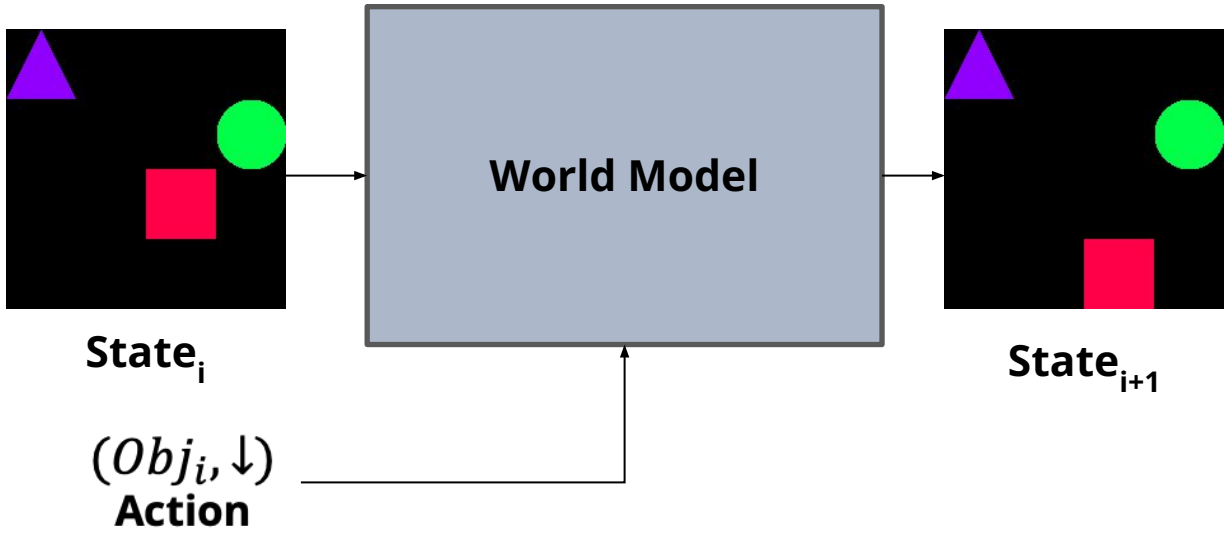
Neurosymbolic Grounding for Object-Oriented Compositional World Models

Atharva Sehgal, Arya Grayeli, Jennifer Sun*, and Swarat Chaudhuri

UT Austin, Caltech*

TL;DW: https://bit.ly/cosmos-wm

World Modeling



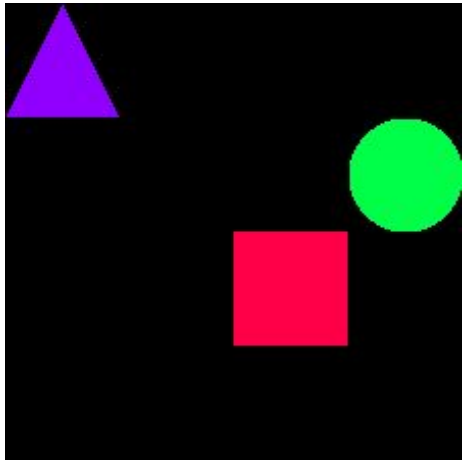
Objective: Implicitly learn the dynamics laws of this domain.

Recurrent world models facilitate policy evaluation. Ha & Schmidhuber, 2018.

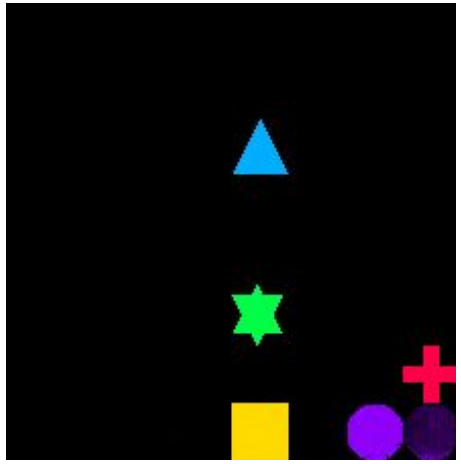
Neural production systems: Learning rule-governed visual dynamics. Goyal, Bengio et al., 2021.

Compositional Generalization

Assuming 5 colors for each object

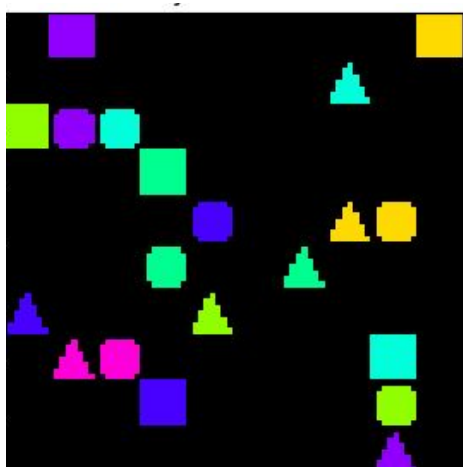


|objects| **= 3*5**
#combinations = **455**



|objects| **= 5*5**
#combinations = **53130**

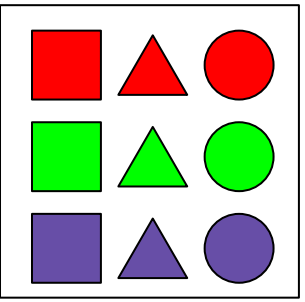
...



|objects| **= 20*5**
#combinations = **5.36×10^{20}**

Compositional Generalization

Distribution of Atoms ($\mathcal{A}$)



Entity Composition

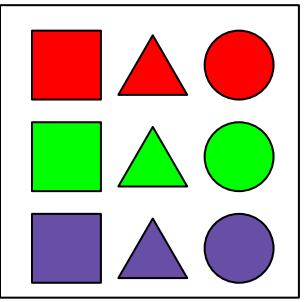
Generalization to composition of shapes not seen together during training

Relational Composition

Entity Composition + Objects with shared attributes have shared dynamics.

Compositional Generalization

Distribution of Atoms ($\mathcal{A}$)



Entity Composition

Generalization to composition of shapes not seen together during training

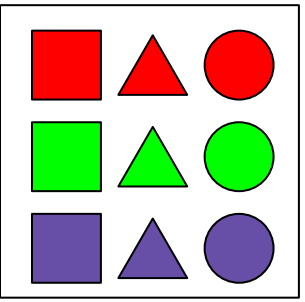
Sampled Composition ($c \sim \mathcal{A}$)

Relational Composition

Entity Composition + Objects with shared attributes have shared dynamics.

Compositional Generalization

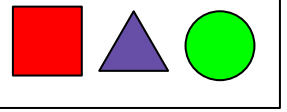
Distribution of Atoms ($\mathcal{A}$)



Entity Composition

Generalization to composition of shapes not seen together during training

Sampled Composition ($c \sim \mathcal{A}$)



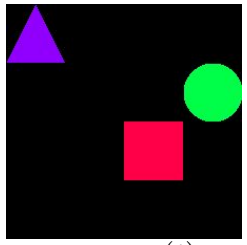
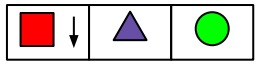


Image ($I^{(t)}$)



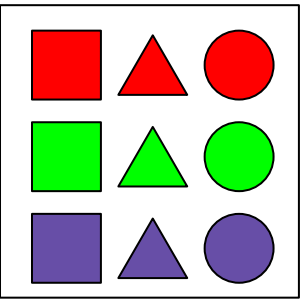
Actions

Relational Composition

Entity Composition + Objects with shared attributes have shared dynamics.

Compositional Generalization

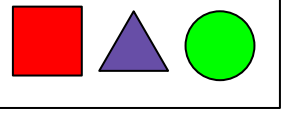
Distribution of Atoms ($\mathcal{A}$)

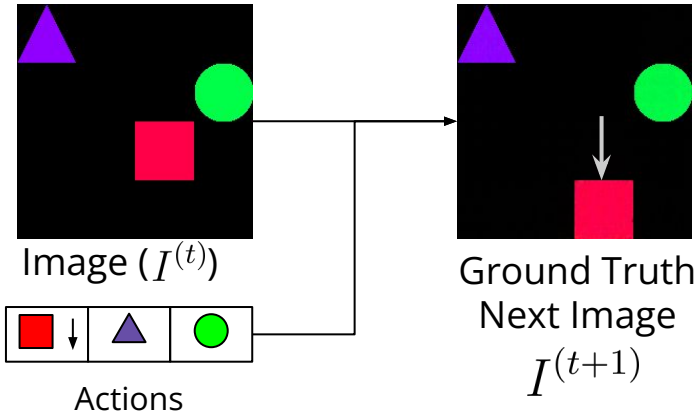


Entity Composition

Generalization to composition of shapes not seen together during training

Sampled Composition ($c \sim \mathcal{A}$)





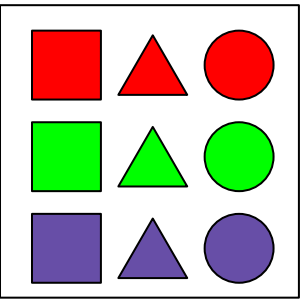
The object dynamics are invariant to distribution shift!

Relational Composition

Entity Composition + Objects with shared attributes have shared dynamics.

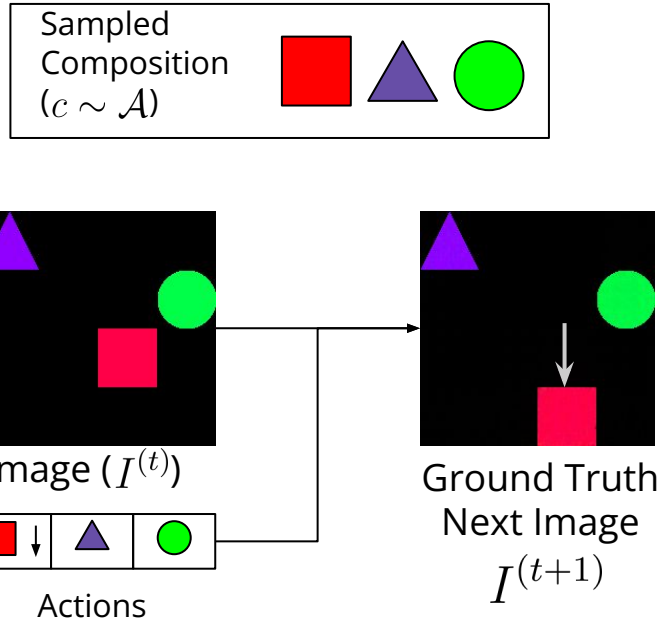
Compositional Generalization

Distribution of Atoms ($\mathcal{A}$)



Entity Composition

Generalization to composition of shapes not seen together during training



The object dynamics are invariant to distribution shift!

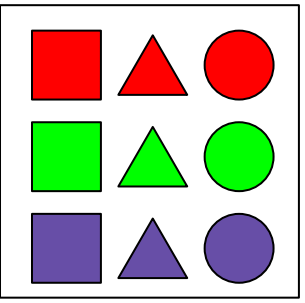
Relational Composition

Entity Composition + Objects with shared attributes have shared dynamics.

Higher level rule: same color => cooperation

Compositional Generalization

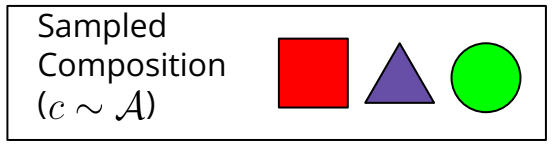
Distribution of Atoms ($\mathcal{A}$)

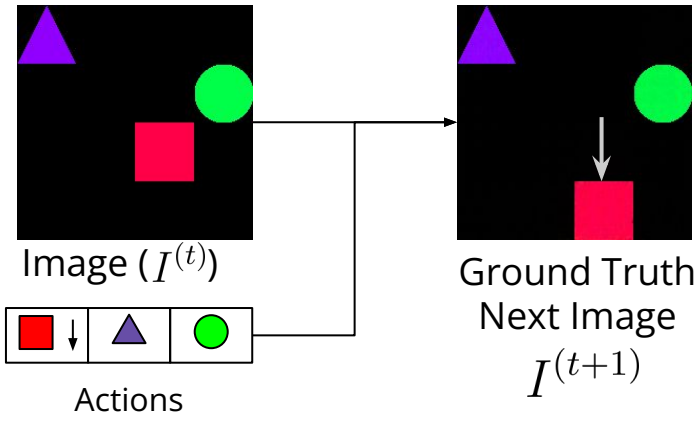


Entity Composition

Generalization to composition of shapes not seen together during training

Sampled Composition ($c \sim \mathcal{A}$)

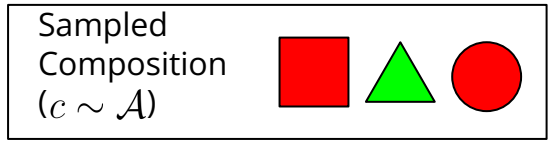




Relational Composition

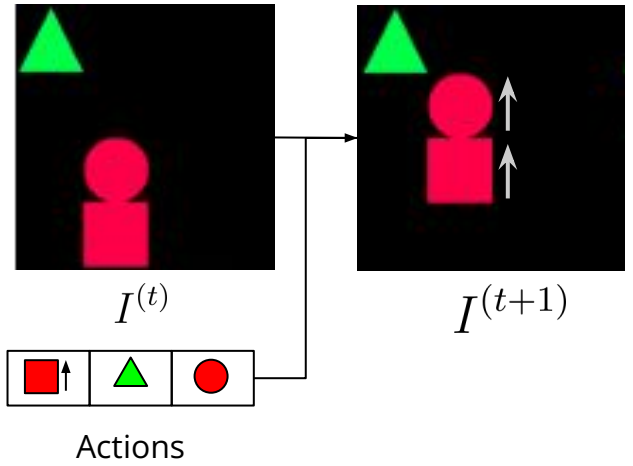
Entity Composition + Objects with shared attributes have shared dynamics.

Sampled Composition ($c \sim \mathcal{A}$)



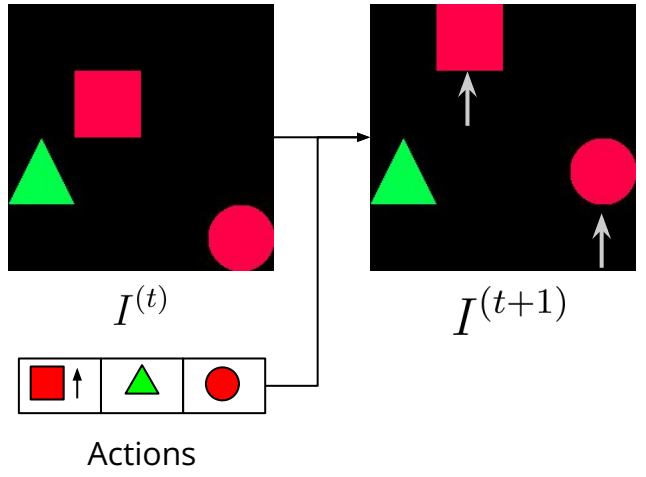
"Sticky"

Shared Attributes = Color + Adjacency

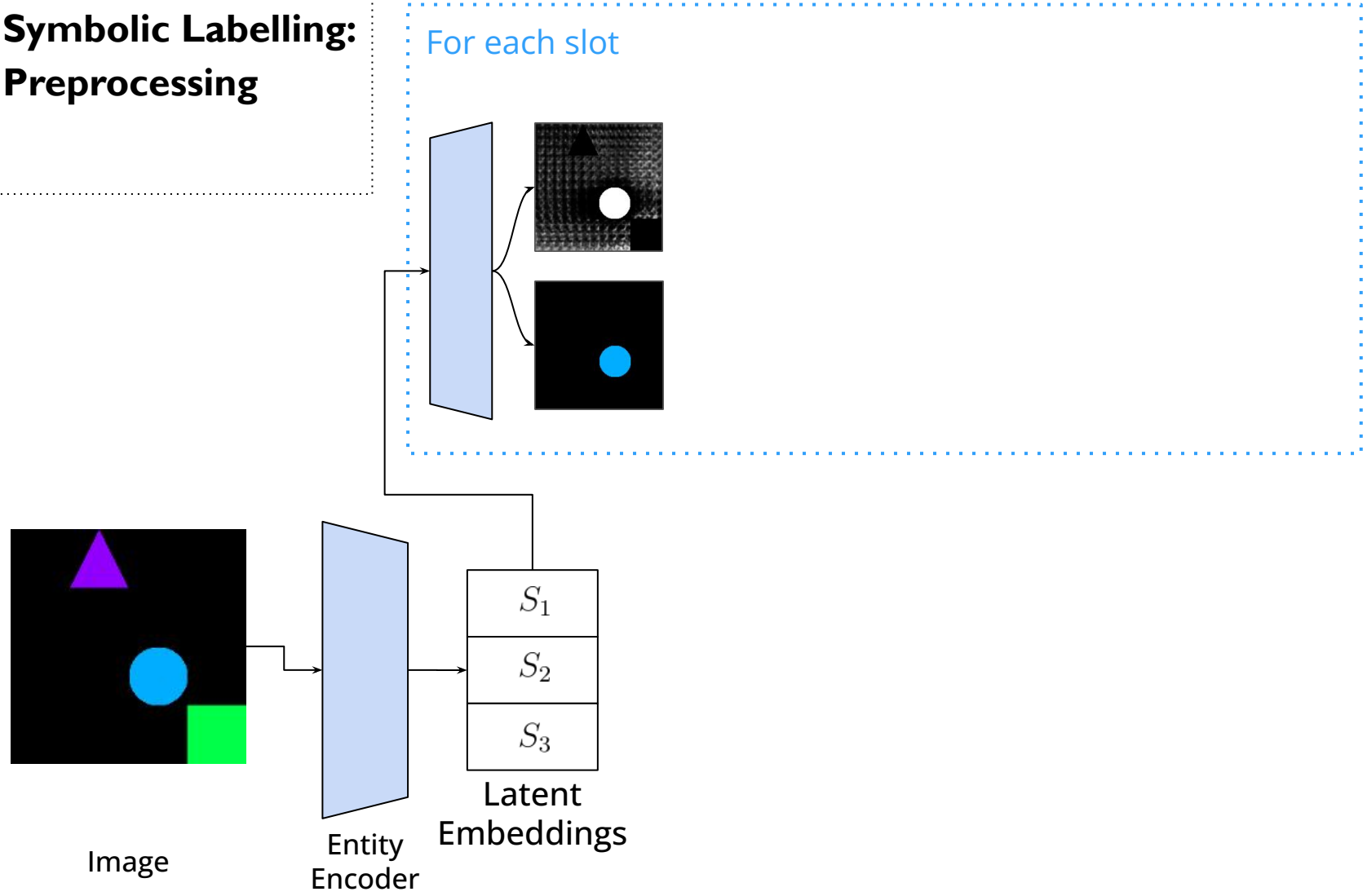


"Team"

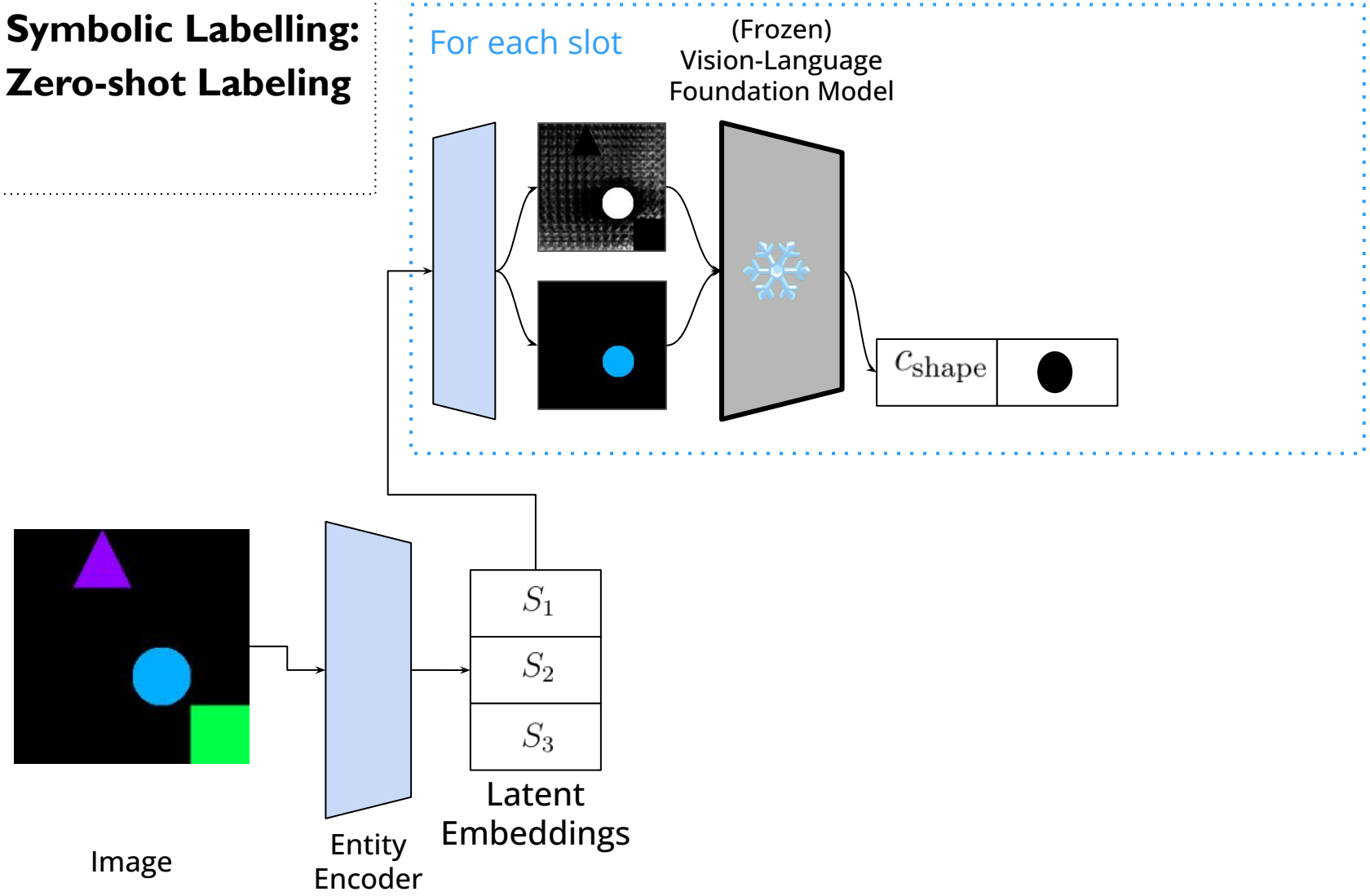
Shared Attributes = Color

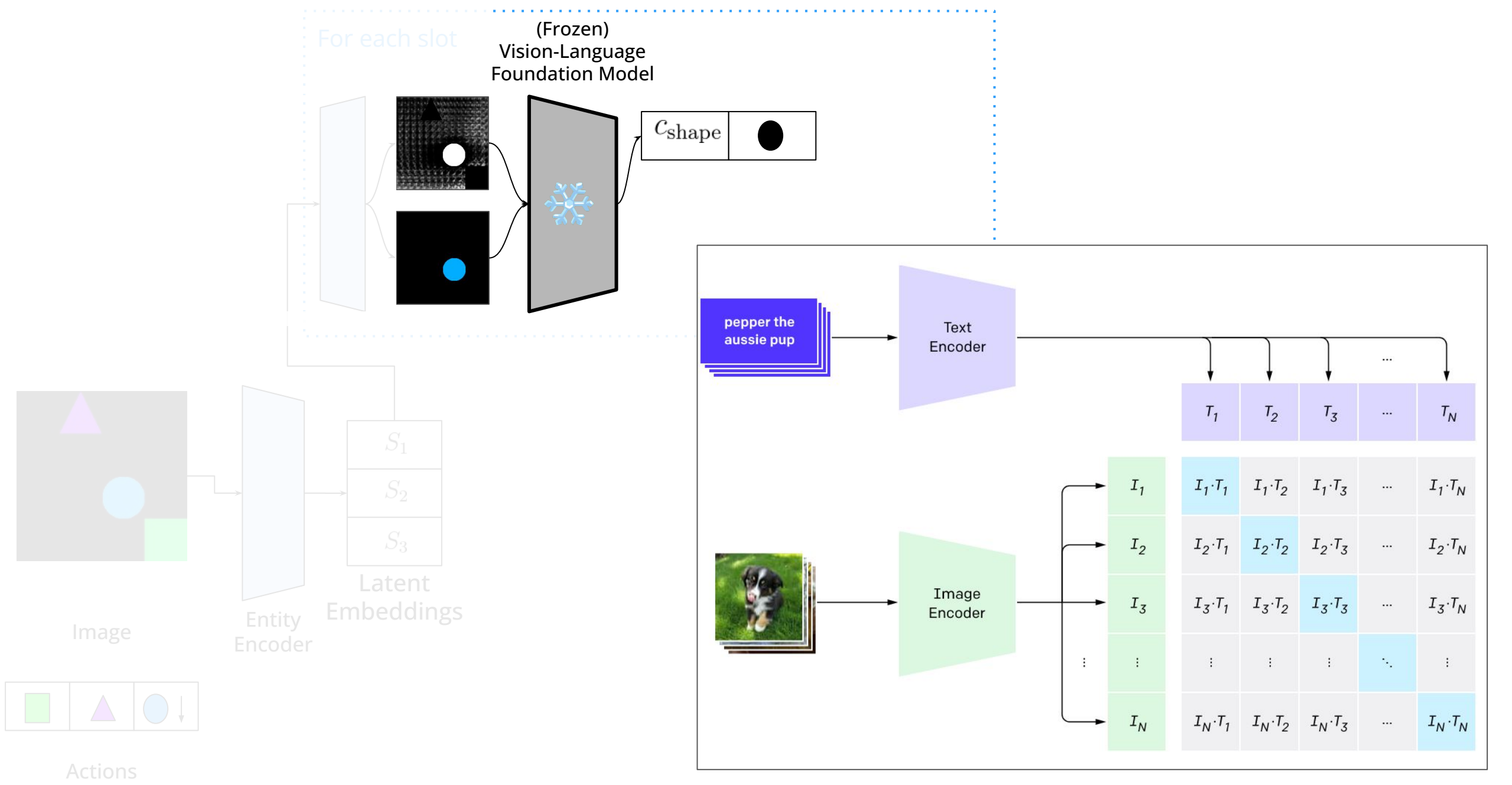


Symbolic Labelling: Preprocessing



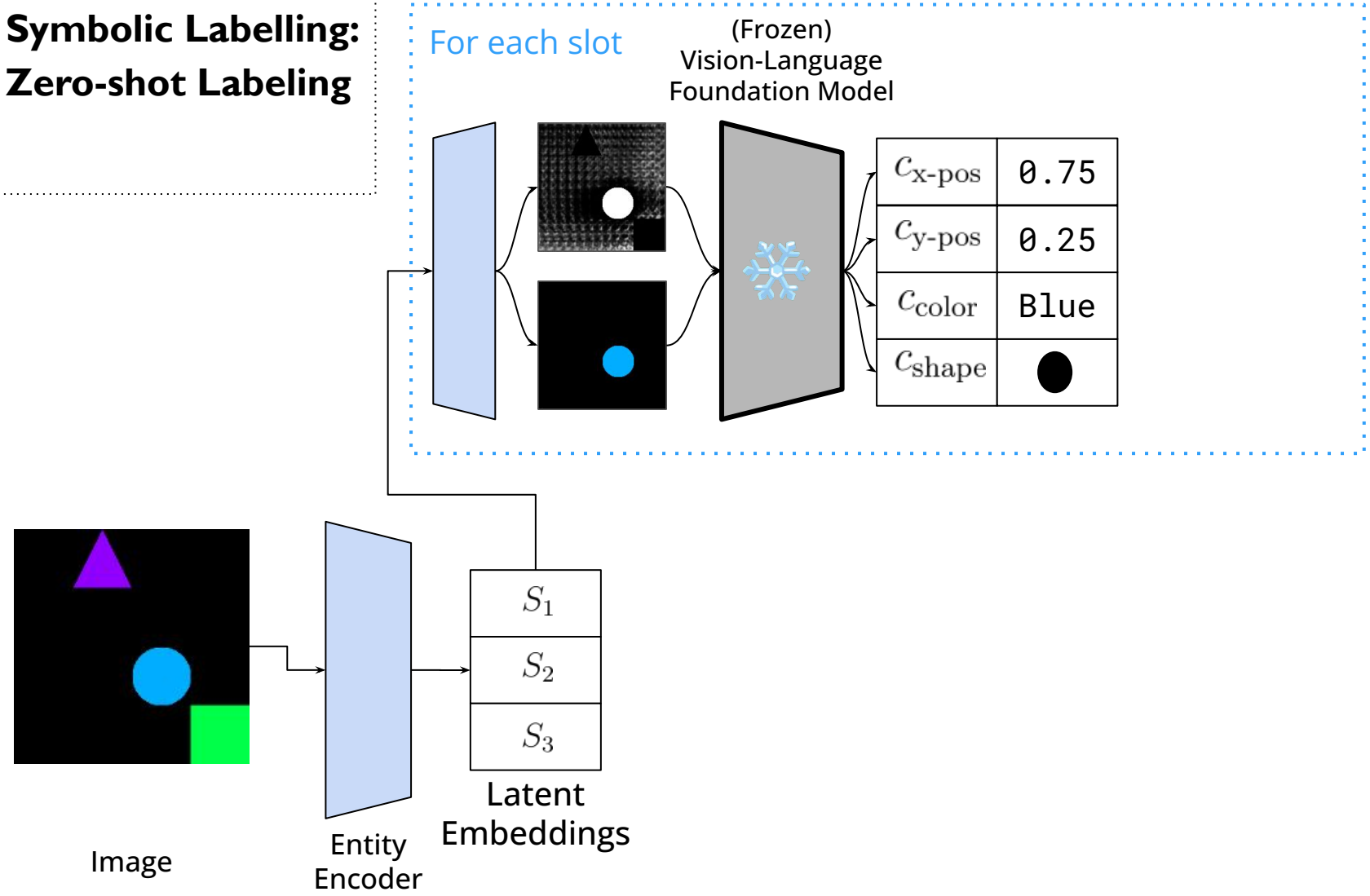
Symbolic Labelling:
Zero-shot Labeling



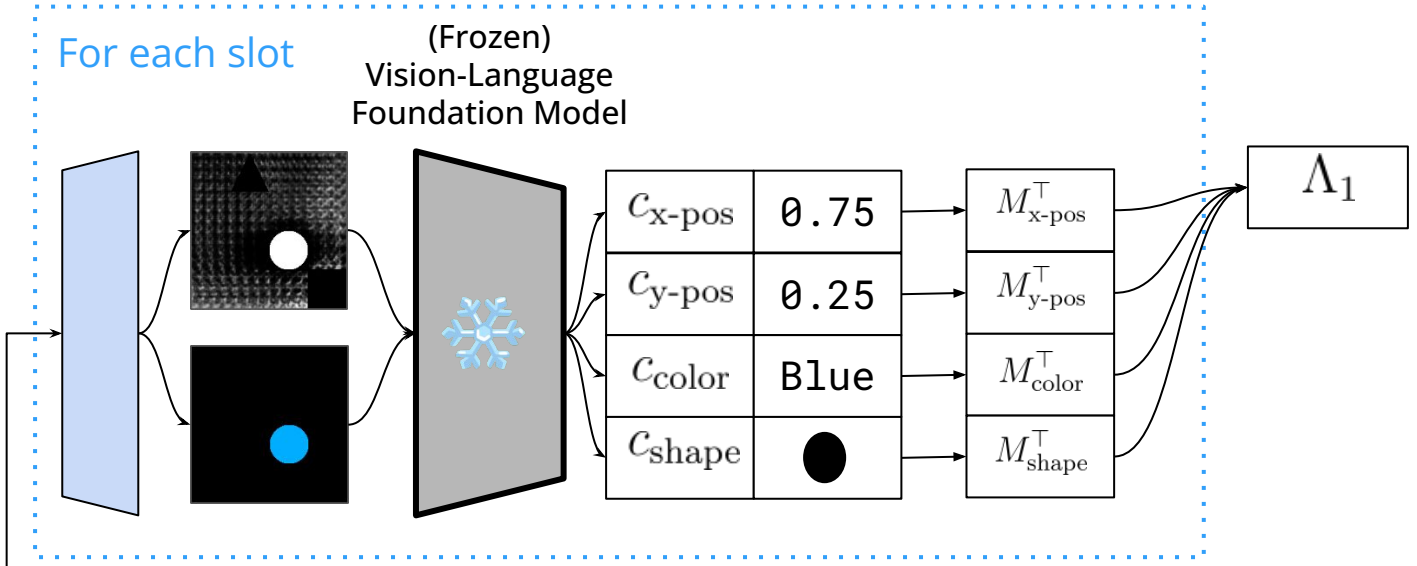


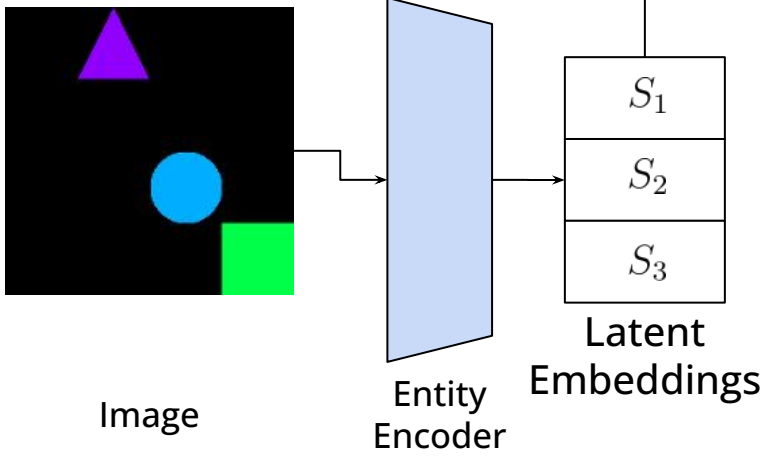
Learning Transferable Visual Models From Natural Language Supervision. Radford et al., 2021.

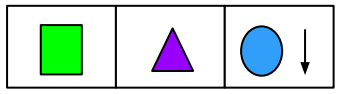
Symbolic Labelling: Zero-shot Labeling



Symbolic Labelling: Concatenation





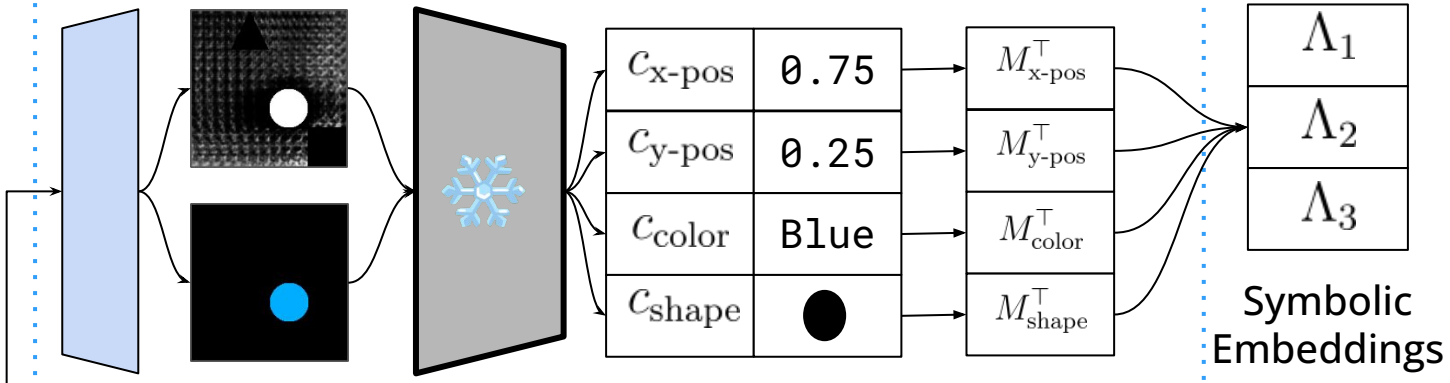


Actions

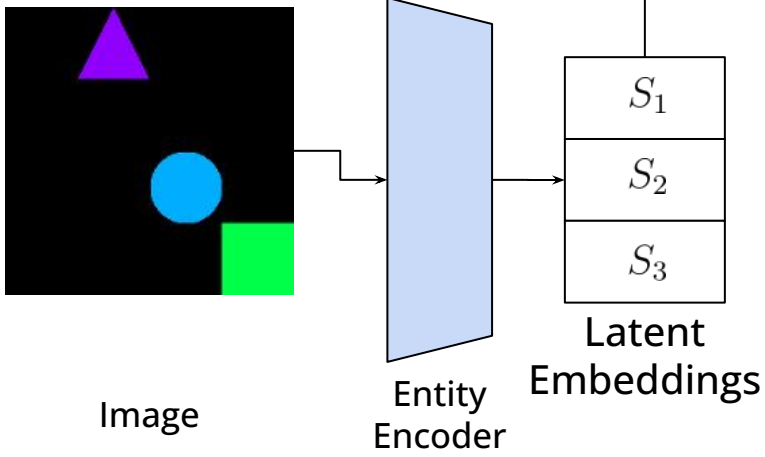
Symbolic Embedding

For each slot

(Frozen) Vision-Language Foundation Model



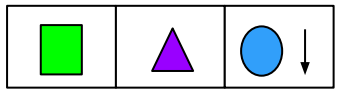
Symbolic Embeddings



Image

Entity Encoder

Latent Embeddings



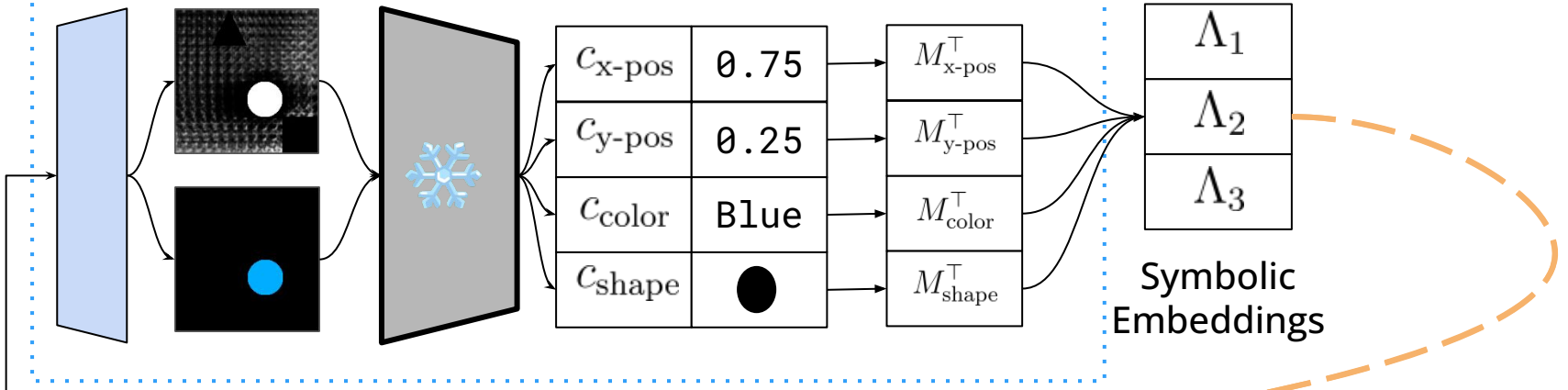
Actions

Strawman A

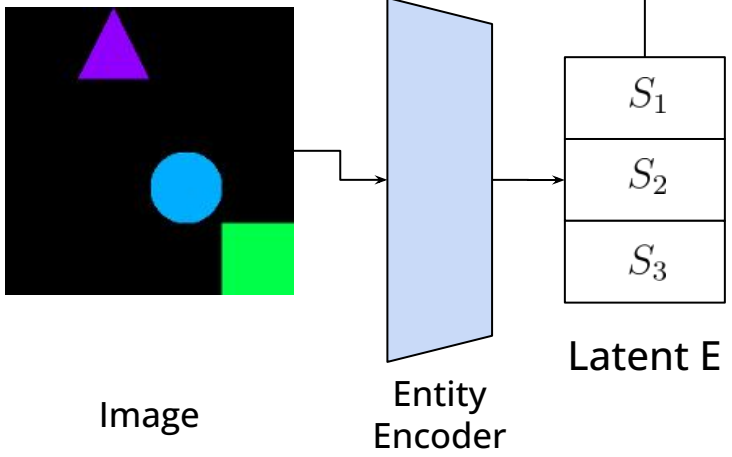
Fully Symbolic Encoding

For each slot

(Frozen) Vision-Language Foundation Model



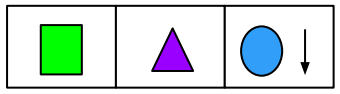
Symbolic Embeddings



Image

Entity Encoder

Latent E



Actions

Selection Module

Execute Module

Decoder

Next Image Prediction

Q: What if we only use Symbolic Embeddings?

✅ : Trivially generalizes to attribute compositions. Robust selection module!

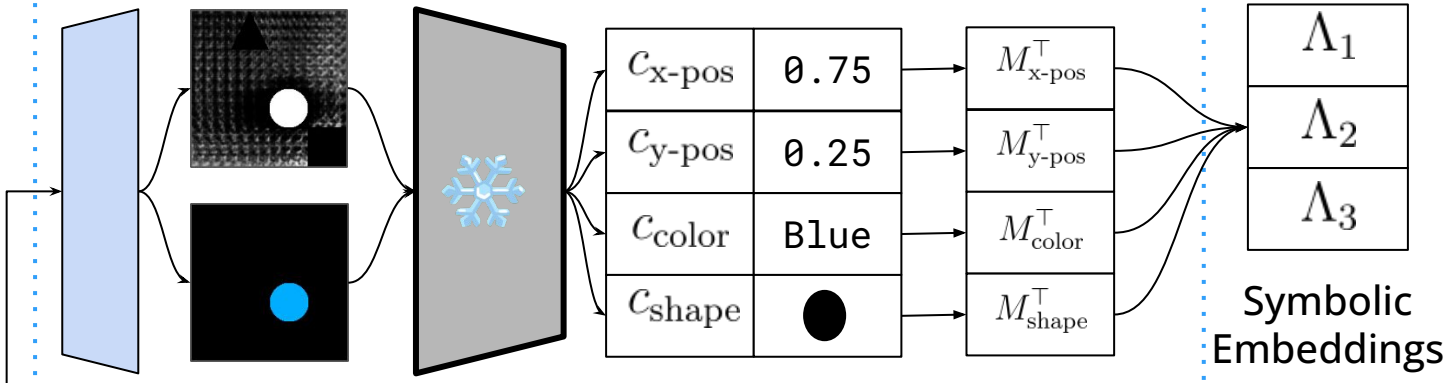
❌ : Symbols bottleneck expressivity. The circle is *Cyan, not Blue!* Erroneous image reconstruction.

Strawman B

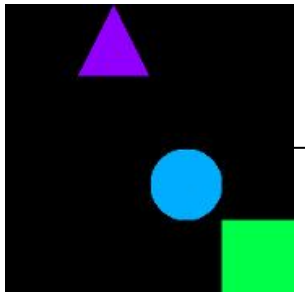
Fully Neural Encoding

For each slot

(Frozen) Vision-Language Foundation Model



Symbolic Embeddings



Image

Entity Encoder

S_1
S_2
S_3

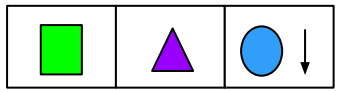
Latent Embeddings

Selection Module

Execute Module

Decoder

Next Image Prediction



Actions

Q: What if we only use Latent Embeddings?

❌ : Tendency to overfit to attribute compositions seen during training.

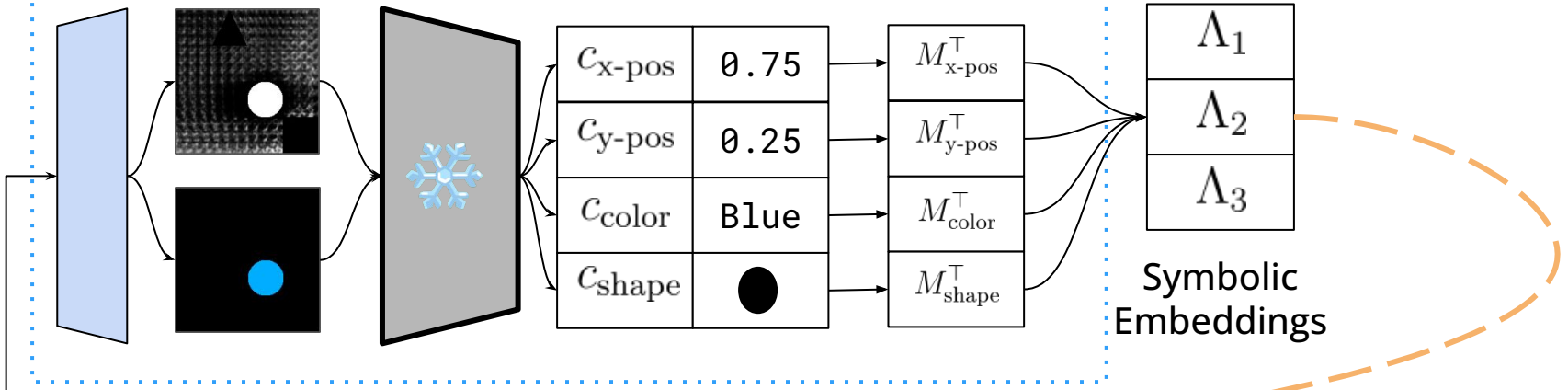
✅ : Latent embedding contains very expressive features. Good image reconstructions!

Hypothesis

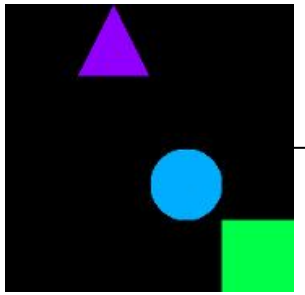
Neuro-symbolic Encoding

For each slot

(Frozen) Vision-Language Foundation Model

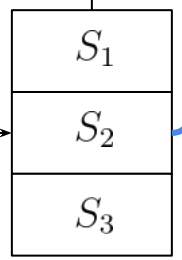


Symbolic Embeddings



Image

Entity Encoder



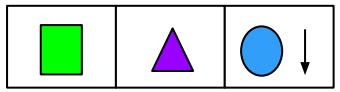
Latent Embeddings

Selection Module

Execute Module

Decoder

Next Image Prediction



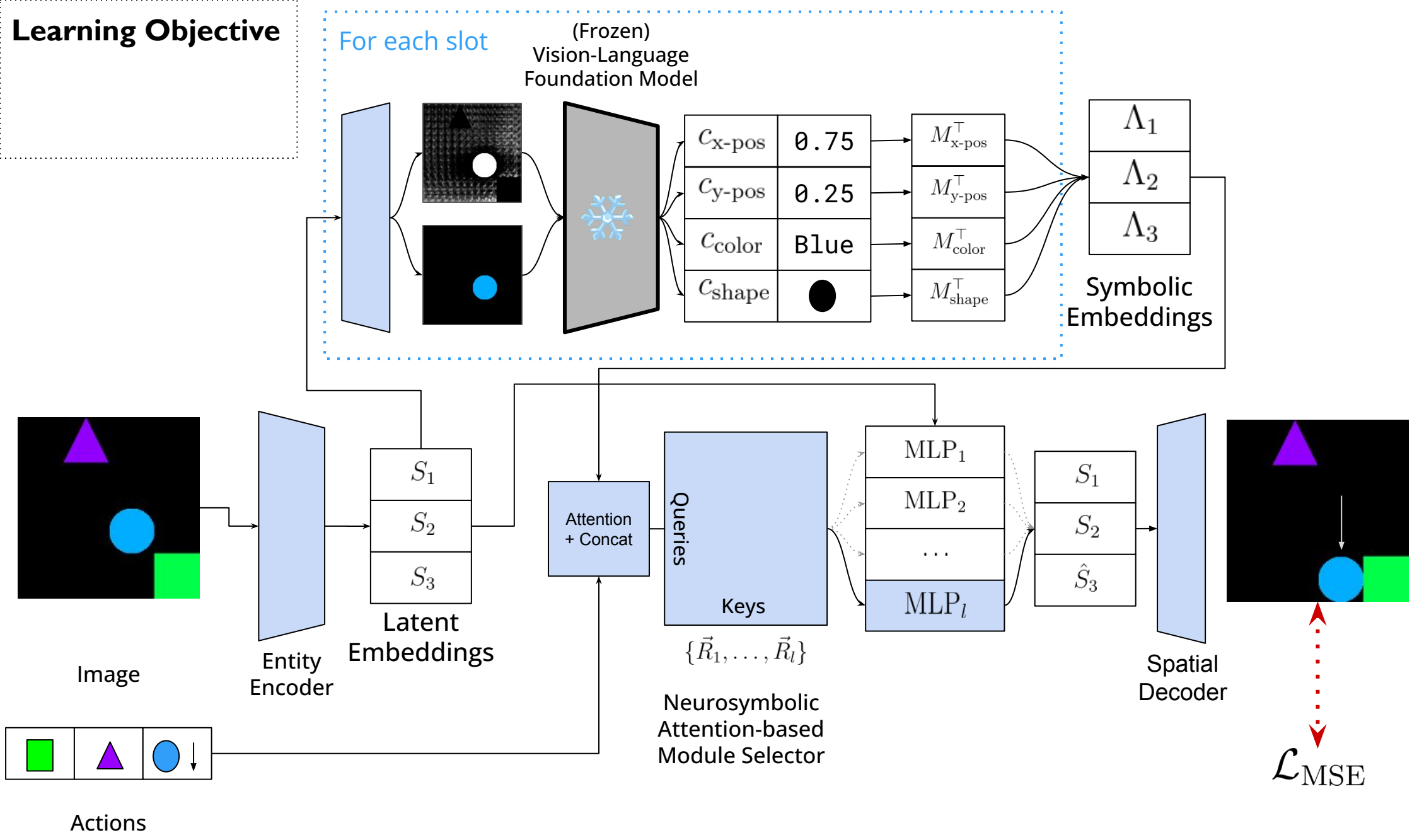
Actions

Using *both* representations:

✅ : Selection robust to attribute compositions.

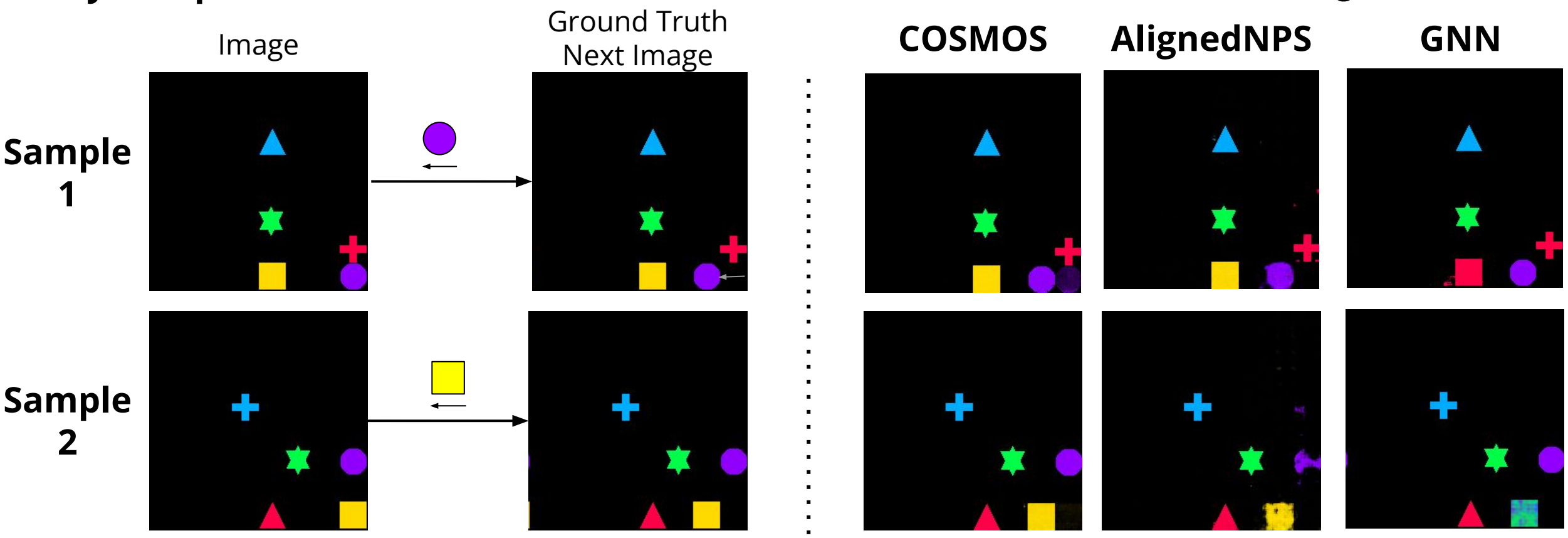
✅ : Latent embedding captures relevant reconstruction features.

Learning Objective



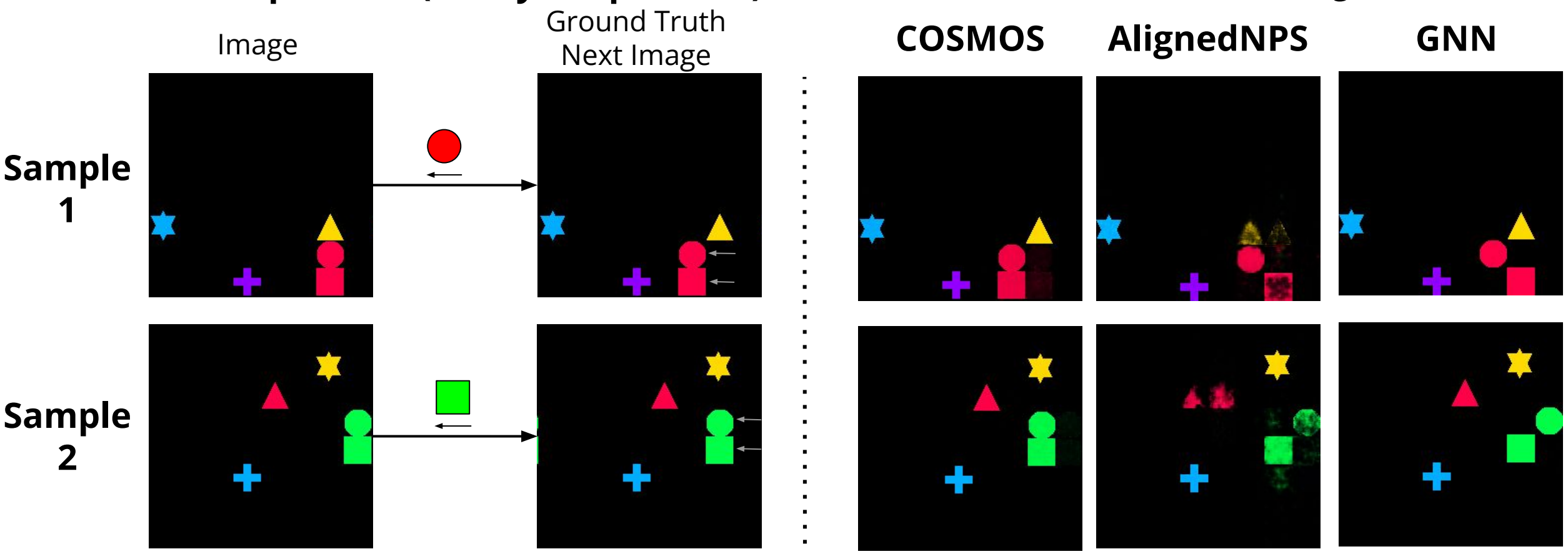
Results

Entity Composition



Results

Relational Composition (Sticky Shapeworld)



Takeaways

- **Explicit symbolic knowledge** helps with compositionality
- **Extend**, rather than replace, deep representations
- **Foundation models** over language (and code) give symbols for free

More Information:

- https://bit.ly/cosmos-wm