

Compositional Learning in OOD Generalization

Srinitish Srinivasan

What do we mean by compositionality?

- A function represented by a function of its parts and by rules that specify how these functions combine.
- Generalization appears when familiar parts support an unfamiliar combination.
- The relevant unit may be a variable, motif, role, or module—not necessarily a word.

Is compositionality binary?

- Probably not. Representations can be more or less decomposable.
- Expressiveness matters: the parts must retain enough information to reconstruct the whole.
- Simplicity matters too: a flexible decoder can hide a complicated code.
- Ask how simply the whole can be described from its constituents.

What kind of OOD case are we asking about?

- Every color is familiar. Every shape is familiar.
- A few color–shape bindings are absent from training.
- The test asks for a new conjunction, not a new constituent.
- So the problem is reuse: can known parts bind without learning the pair?

Why is binding the hard part?

- Recognizing red and circle separately is not enough.
- The representation must preserve which attribute belongs to which object.
- CLIP can retain object information while its binding function remains combination-specific.

Does more data solve this?

- Not by volume alone.
- More examples from the same combinations may leave novel-combination performance unchanged.
- Broader combinatorial coverage encourages more factorized representations.
- Pretrained models show partial structure, but diverse combinations still matter.

What exactly is held out?

- Training uses 40,000 two-object scenes.
- The primitives are five colors and five shapes.
- Twenty color–shape cells are observed. Diagonal cells are held out.
- IID validation uses observed pairs. OOD validation contains held-out pairs.

What do the Gaussian slots capture?

- A ViT maps the image to patch tokens.
- Learned queries pool different evidence into slots: color, shape, position, or other object cues.
- Each slot stores a mean and uncertainty rather than a single point.
- Slot meaning is learned from reconstruction; an index has no permanent semantic label.
- Uncertainty lets confident dimensions contribute more strongly later.

How are slots composed?

- A sparse router selects the slots relevant to each higher-level representation.
- A product of experts combines compatible evidence and weighs confident dimensions more.
- The composed representation is passed through the frozen decoder for reconstruction.
- At test time, the same rule can combine familiar evidence in a held-out pairing.
- Reconstruction is the training signal; the composition step itself is fixed.

Where could the OOD gain come from?

- The model has seen red in other combinations.
- It has seen circle in other combinations.
- If their latent directions are reusable, routing can select both at test time.
- The same product can then form the absent red-circle conjunction.

So what do we take away?

- Compositional OOD is a binding problem: familiar parts, unfamiliar pairing.
- Sparse routing and a product of experts provide an explicit composition rule.
- The claim is empirical: reusable factors and routes still have to emerge.
- The next test is whether the mechanism survives beyond controlled reconstruction.