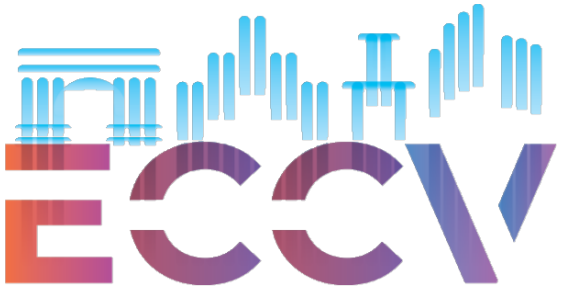
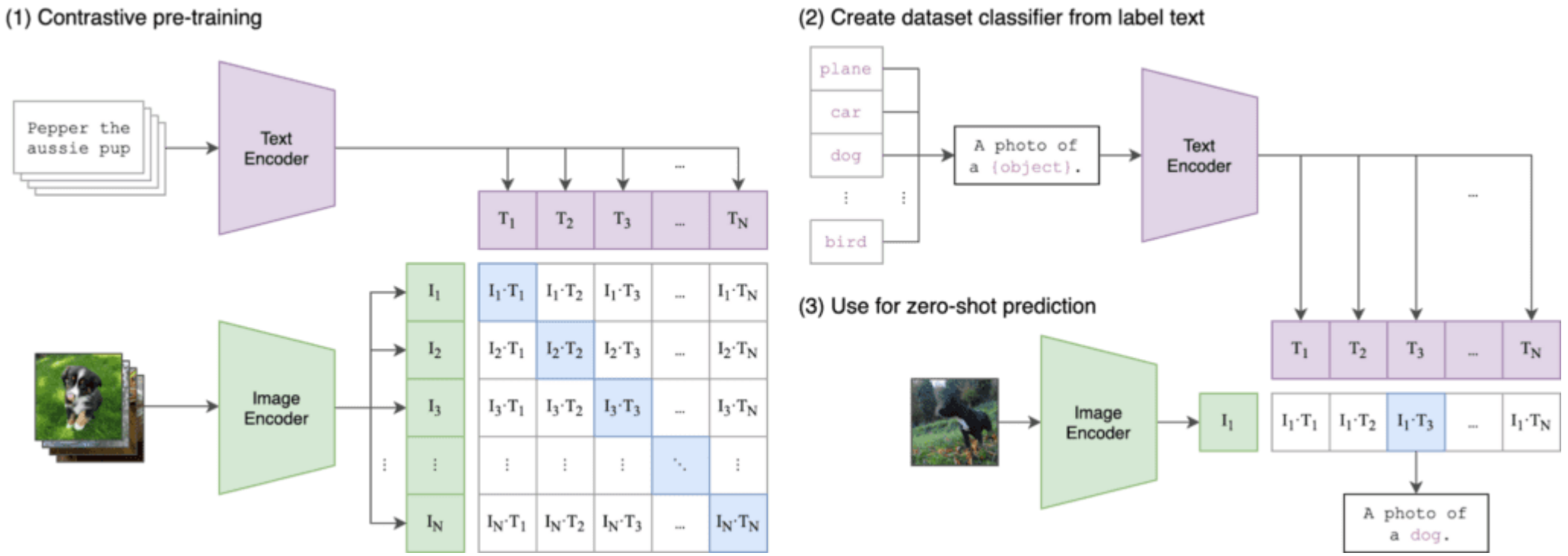


Deciphering the Role of Representation Disentanglement: Investigating Compositional Generalization in CLIP Models

Reza Abbasi, Mohammad Hossein Rohban, Mahdieh Soleymani Baghshah

CLIP

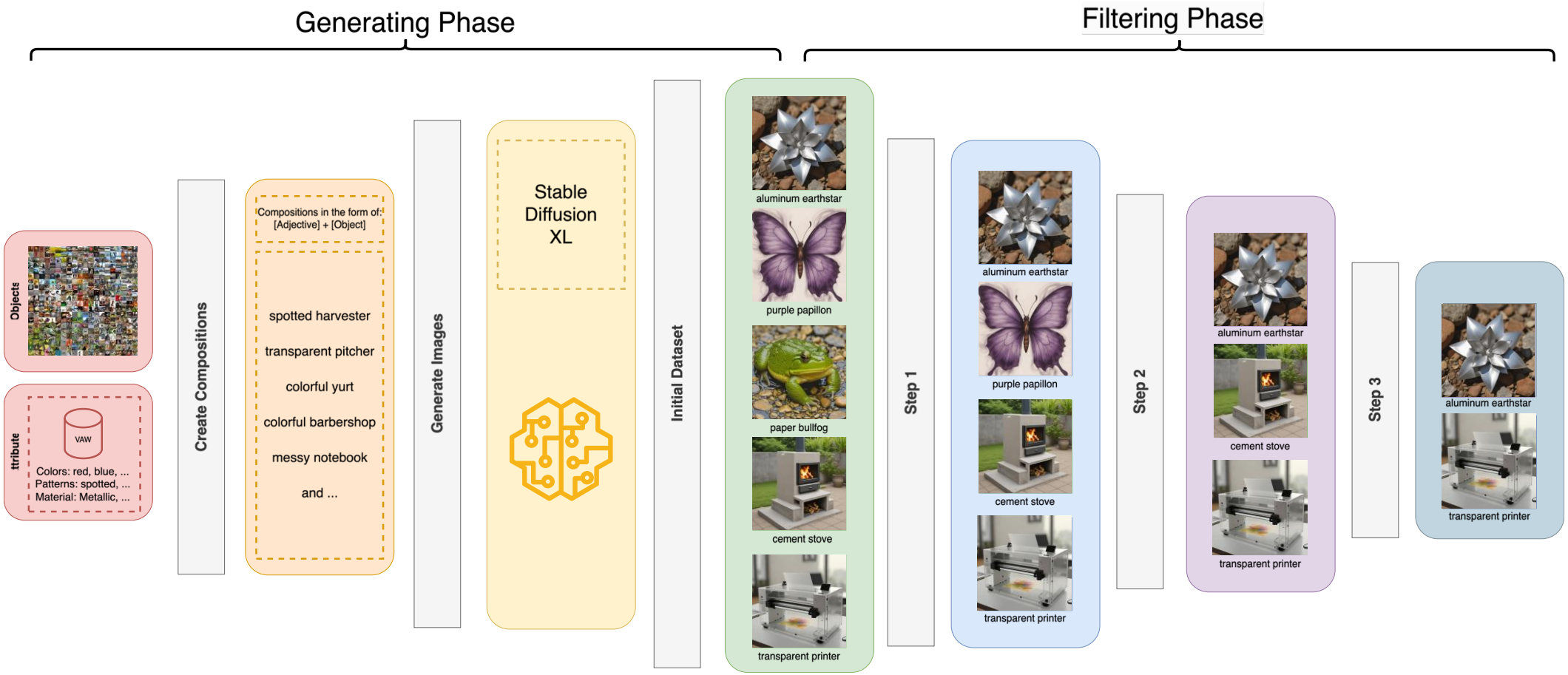

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

CLIP

- CLIP models show Out of Distribution (OoD) generalization
- Compositional OoD (C-OoD) generalization is unexplored
- We investigate factors contributing to C-OoD in CLIPs

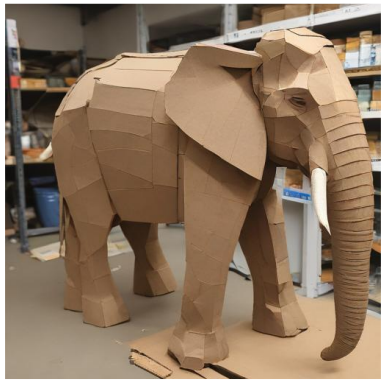
Existing datasets fail to provide true C-OoD scenarios for CLIP

ImageNet-AO Dataset



ImageNet-AO Dataset

- Image Generation
  - Select objects and attributes list from existing Datasets
  - Generate Images by SDXL-turbo
- Filtering Process
  - Exclusion of Known Combinations



cardboard tusker

cement bell pepper

checkered platypus

straw bathtub

chocolate beaver

curly carton

dilapidated hog

colorful solar dish

cobblestone trailer

fluffy warplane

red badger

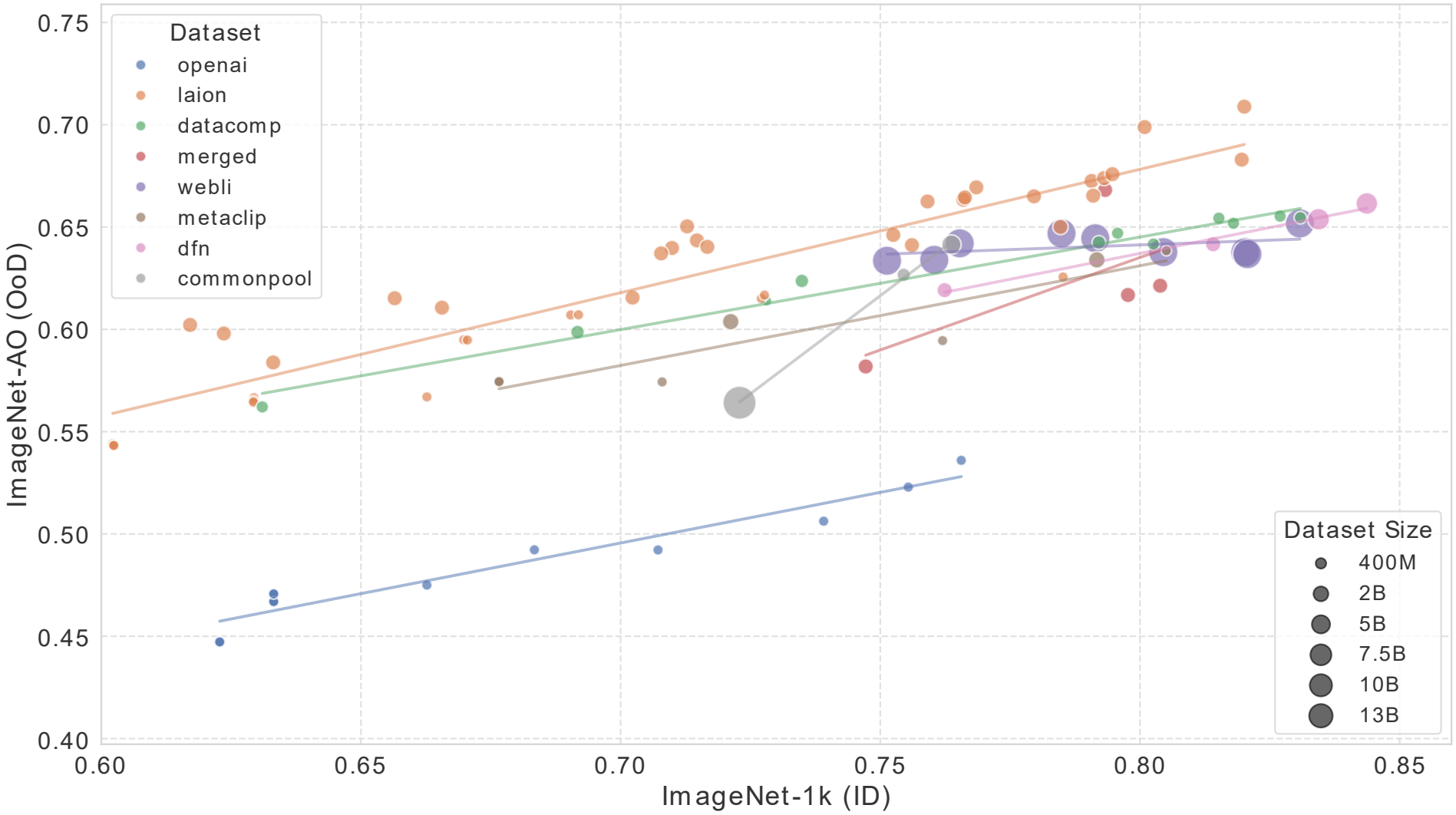
paper caldron

transparent cash machine

fabric police van

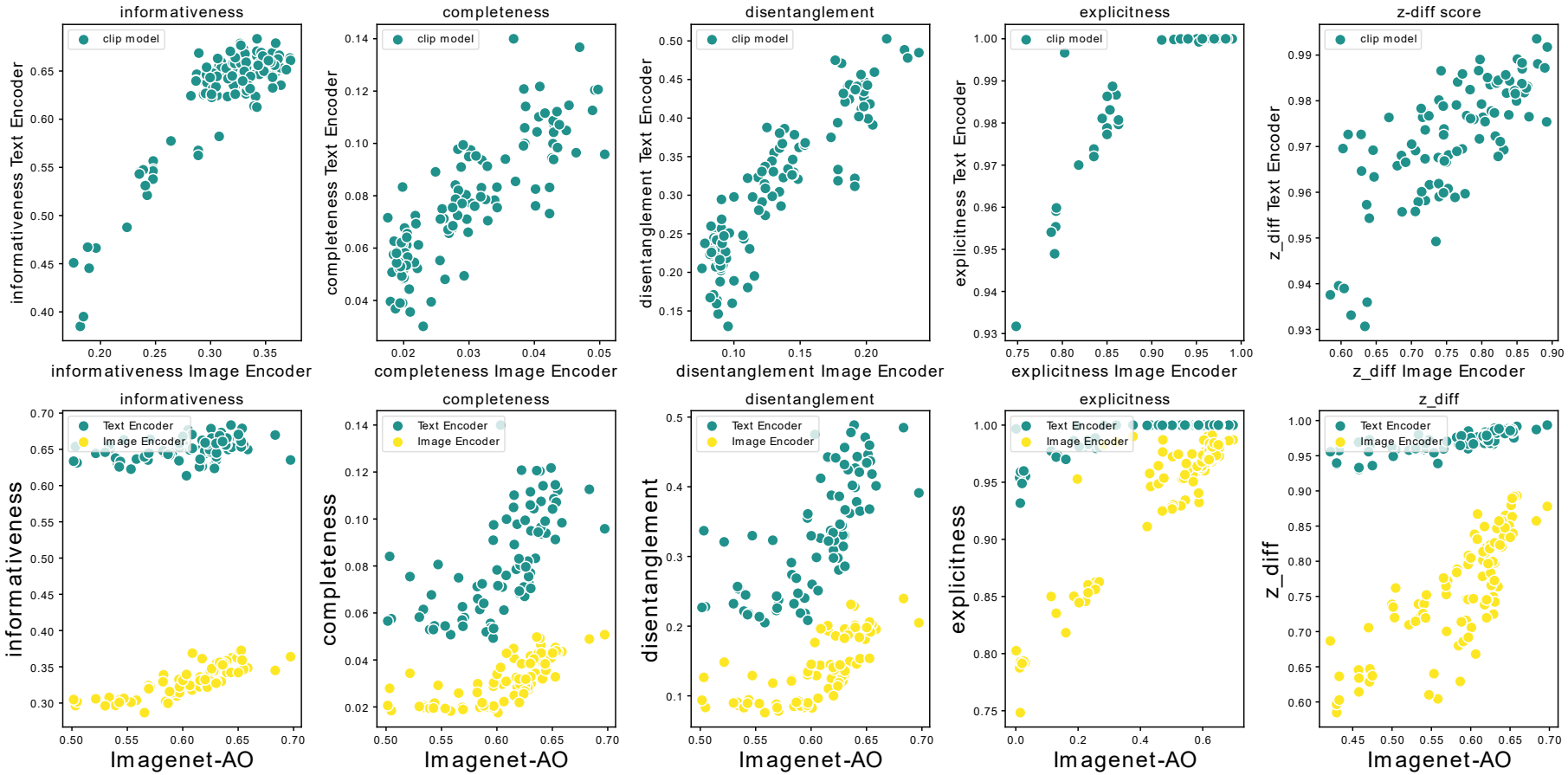
purple crayfish

Comparison of CLIP Models on ImageNet-AO

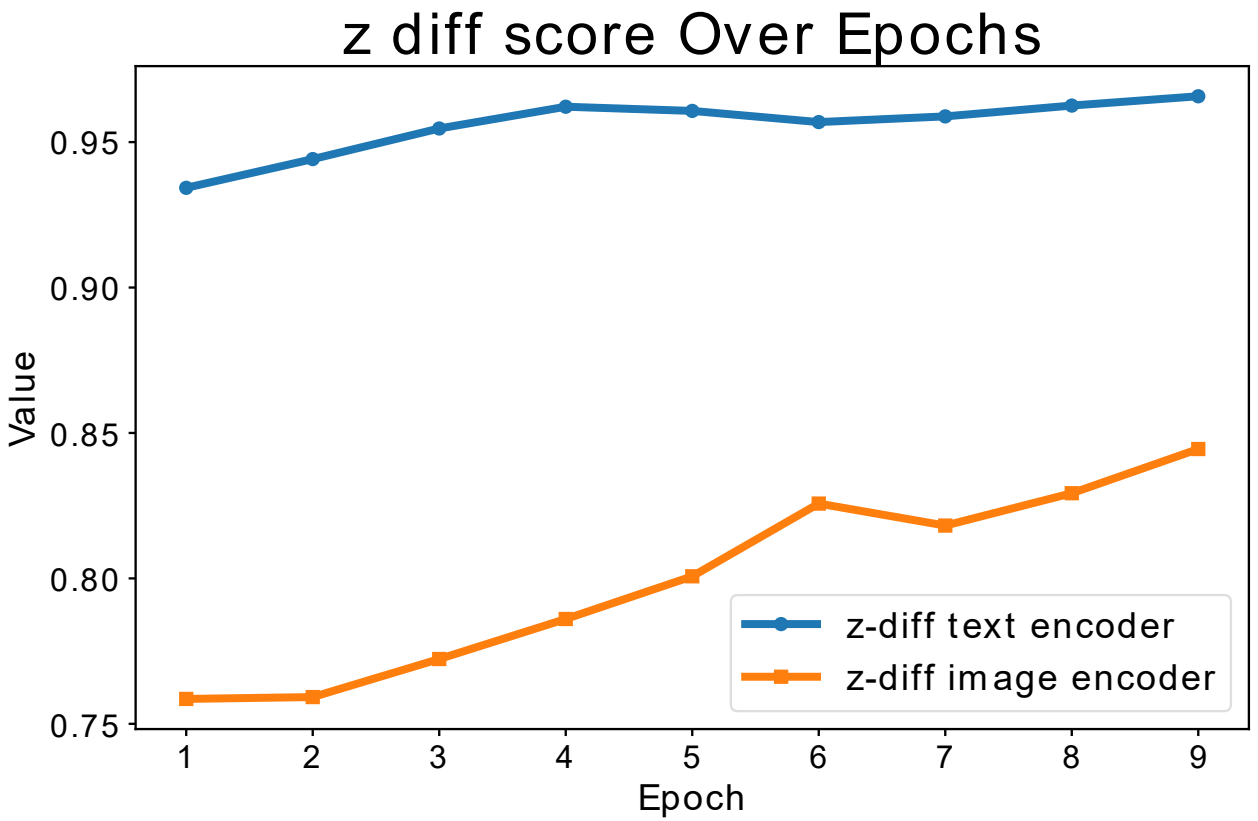


Why CLIP has Compositional Generalization?

Disentanglement



Track embedding disentanglement during training



Thank you for Watching