# Decomposing and Interpreting Image Representations via Text in ViTs Beyond CLIP

Sriram Balasubramanian, Samyadeep Basu, Soheil Feizi

**How to automatically interpret components of arbitrary ViTs using a CLIP text encoder**

## Introduction

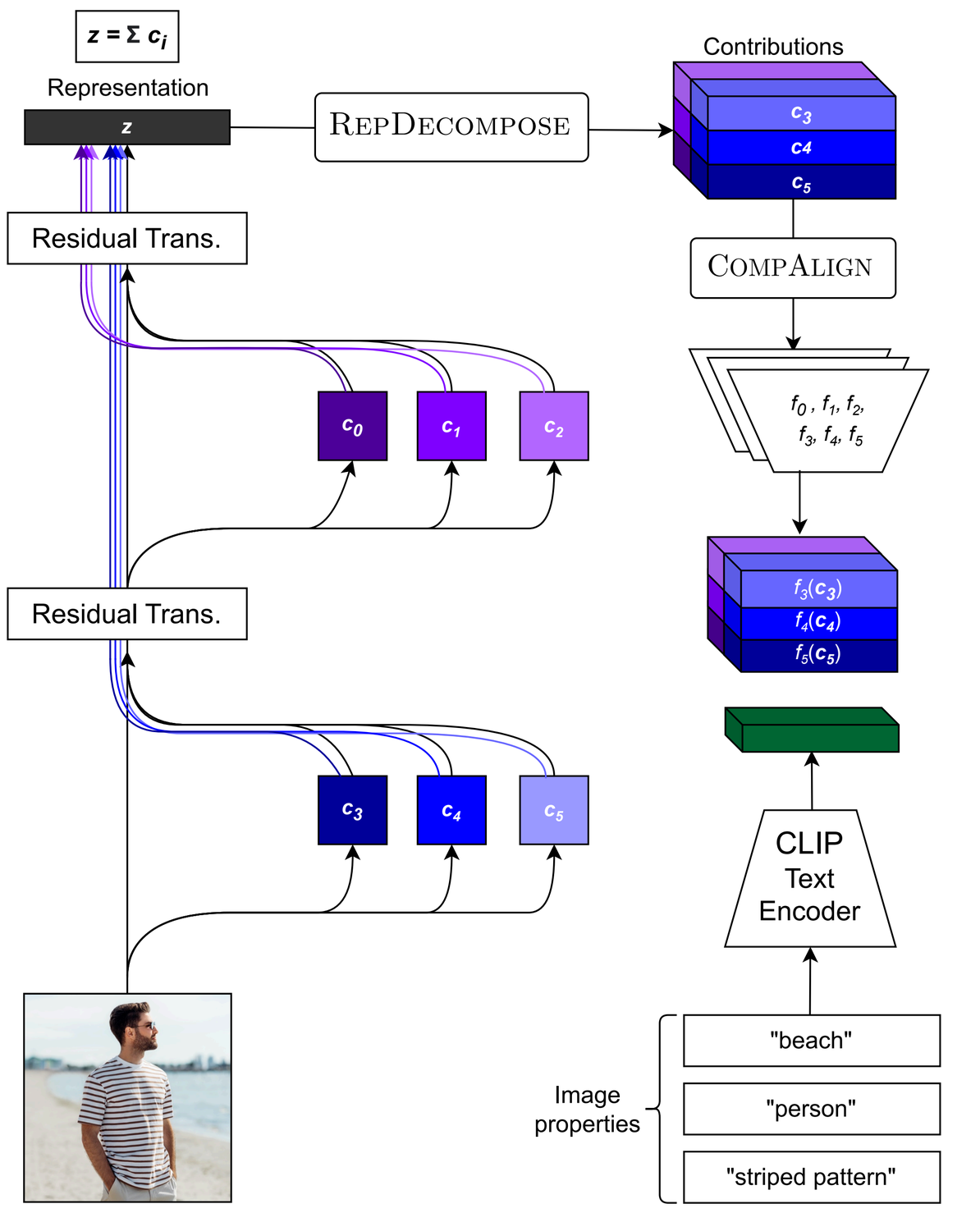
Recent works have explored how individual components contribute to the final image representation of CLIP-ViT by leveraging the shared image-text representation of CLIP. Contributions of each individual component can be interpreted via text using the CLIP text encoder. However, there are challenges in extending this to arbitrary ViTs:

1. Lack of corresponding text encoder means that component contributions cannot be interpreted via text.
2. Significant manual effort is required to reconstruct these contributions for each model architecture as they are often not explicitly computed.

We thus introduce **CompAlign**, a method to align the contributions from each component to CLIP space to enable text-based interpretation, as well as **RepDecompose,** an algorithm which traverse the model's computational graph to extract contributions in an architecture-agnostic manner.

In general, there is no one-to-one mapping between components and image feature, therefore we introduce a **scoring function** to assign the importance of a feature to a given component and vice versa. Using this, we can perform tasks like image retrieval, token contribution visualization, and spurious correlation mitigation by carefully selecting or ablating specific components.

## Method



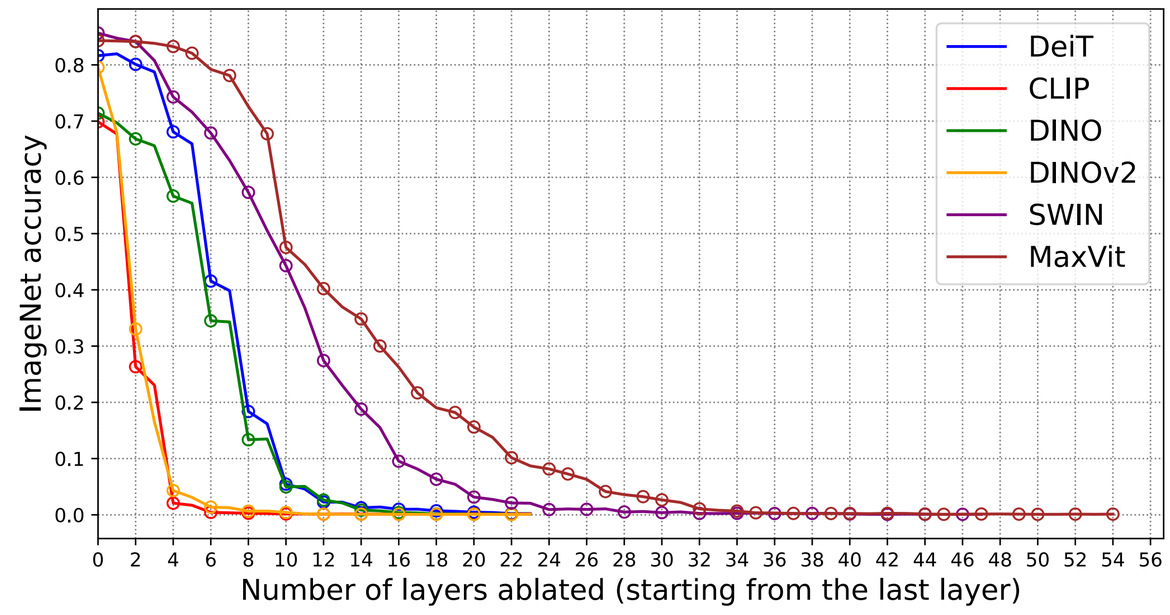

### CompAlign

$$L(\{f_i\}_{i=1}^N) = \mathbb{E}_{\{\mathbf{c}_i\}_{i=1}^N, \mathbf{z}_{\text{CLIP}}} \left[ 1 - \cos\left( \sum_i f_i(\{\mathbf{c}_i\}), \mathbf{z}_{\text{CLIP}} \right) \right] + \lambda \sum_i \| f_i^T f_i - I \|_F$$

Scoring function which computes correlation of component contributions C with the final representation Z along a feature basis B

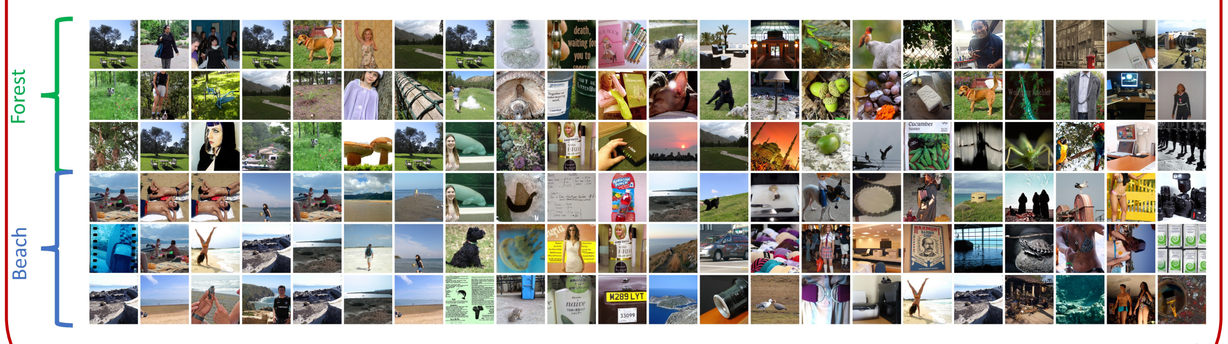
**function** CompAttribute($C$, $Z$, $B$)
- $B \leftarrow \text{orthogonalize}(B)$
- $s_Z \leftarrow ZB^\top$
- $s_C \leftarrow CB^\top$
- $r \leftarrow \text{correlation\_coefficient}(s_Z, s_C, \text{dim=0})$
- **return** mean($r$)

## Layer-wise Ablation

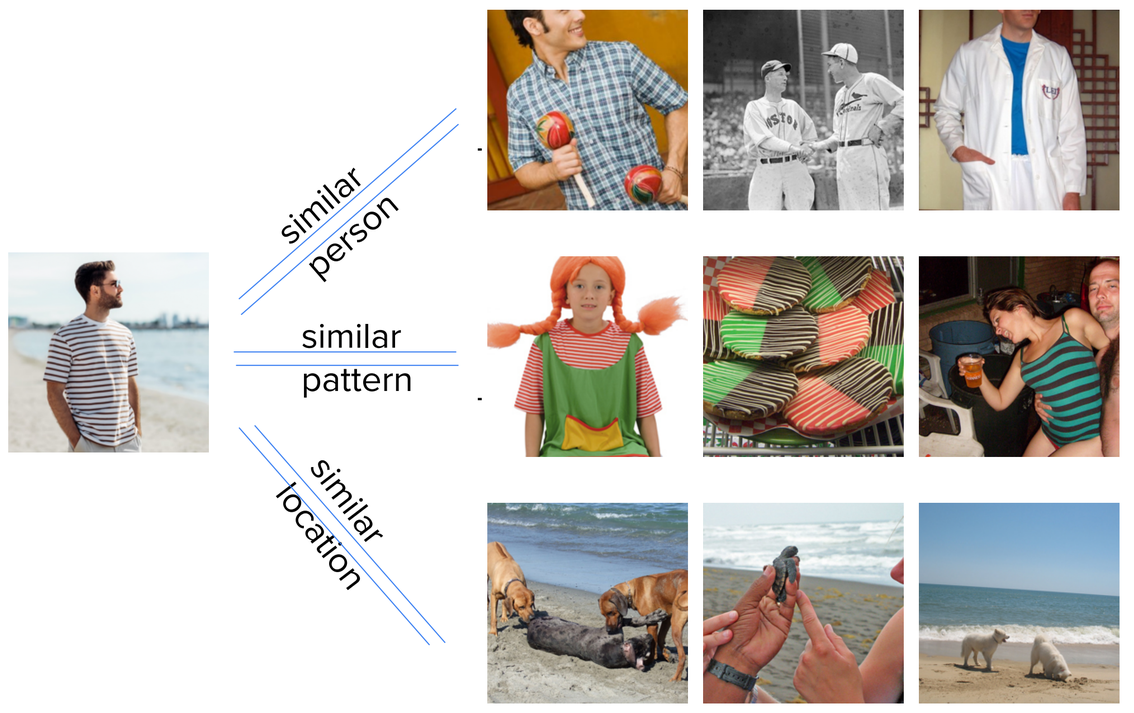



## Practical Applications

### Text-based Image Retrieval




### Image-based Image Retrieval



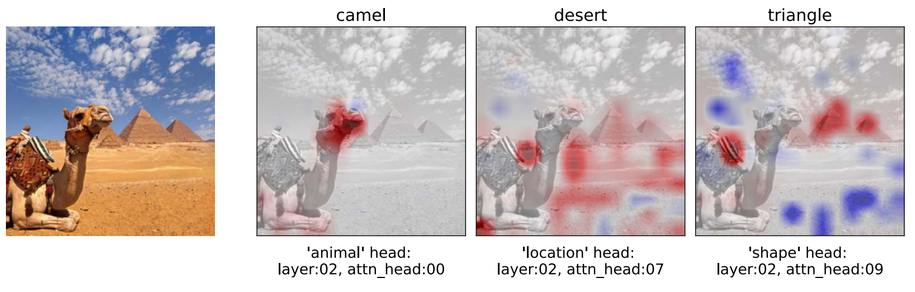

### Spurious correlation mitigation

| Model name | Worst group accuracy | Average group accuracy |
|---|---|---|
| DeiT | 0.733 → **0.815** | 0.874 → **0.913** |
| CLIP | 0.507 → **0.744** | 0.727 → **0.790** |
| DINO | 0.800 → **0.911** | 0.900 → **0.938** |
| DINOv2 | 0.967 → **0.978** | 0.983 → **0.986** |
| SWIN | 0.834 → **0.871** | 0.927 → **0.944** |
| MaxVit | 0.777 → **0.814** | 0.875 → **0.887** |

Worst group accuracy and average group accuracy for Waterbirds dataset before and after intervention for various models

(Format: before ➔ after)

### Visualization of Token-Component contributions




**PAPER**