# Core VLM Components

## ViT

- ViT stands for Vision Transformer — a model architecture introduced by Google Research in 2020 that applies the transformer architecture (originally designed for NLP) directly to image data.
- Outperforms CNNs on many vision tasks and is used in many modern VLMs as the image encoder.
- Forms the backbone of models like CLIP, DINOv2, and Google Lens.