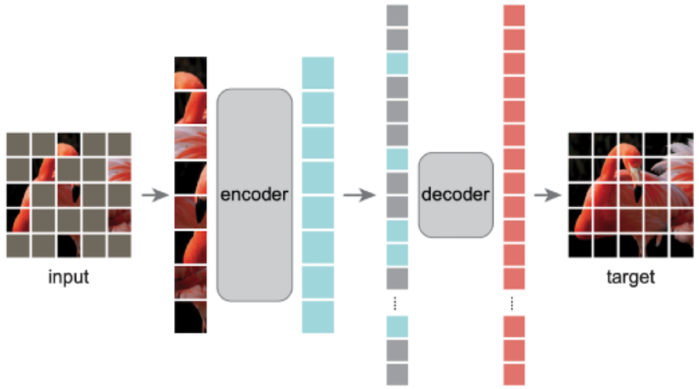
Design: *built upon cutting-edge ML methods such as MAE & Vision Transformer*

Masked Auto-encoders

Learning with fill in the blanks

The "How":

- *Mask:* A large portion of the input (e.g., 75% of patches)
- *Encode:* A deep Encoder processes *only the visible patches*.
- Reconstruct: A Decoder guess the missing patches.



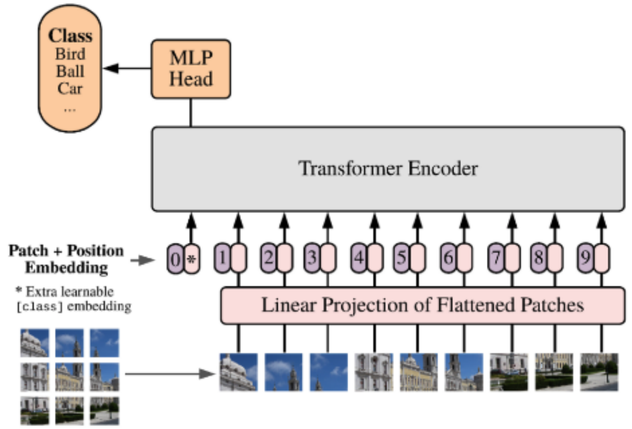
- **The Goal:** force Encoder to learn a rich representation of the data, not just surface-level details.

Vision Transformer

Self-attention to access global features

The "How":

- *Patchify*: An image is broken down into a sequence of patches.
- *Embed*: Each patch is converted to feature vector + positional info
- *Transformer Encoder*: self-attention to model the token relation

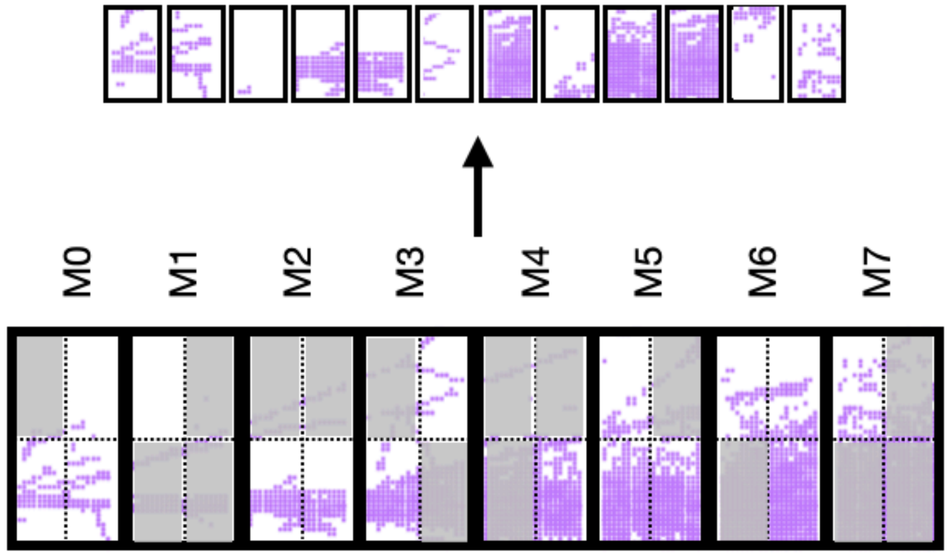


- **The Goal:** capture long-range dependencies and global context across the entire input.

Self-Supervised?

Pipeline

- What does it mean self-supervised?
  - A type of AI that learns from large amounts of unlabeled data by creating its own "labels" or "supervision" from the structure of the data itself - by predicting missing or altered parts of the input



Info on the masked part?: Occupancy, not the Energy