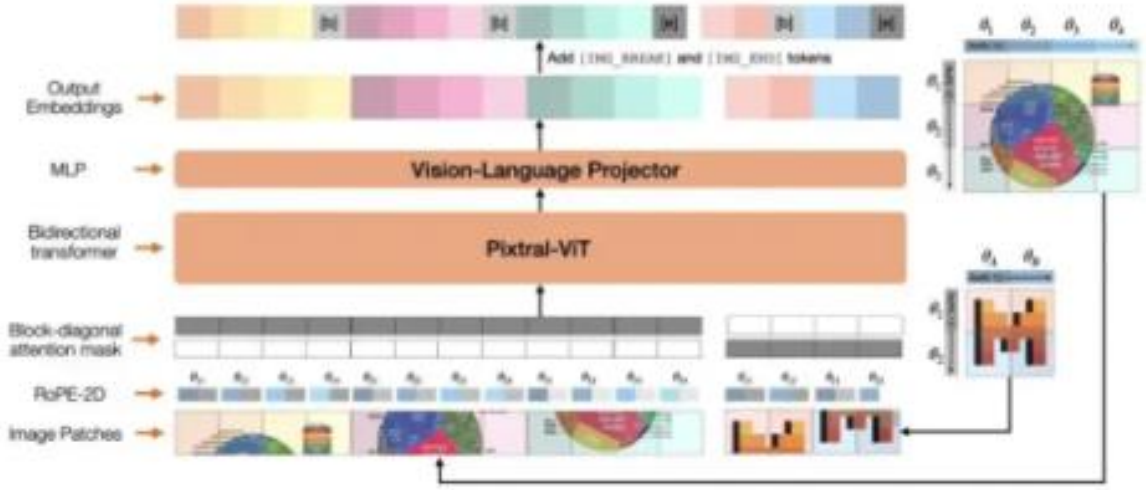

Figure 2: Pixtral Vision Encoder. Pixtral uses a new vision encoder, which is trained from scratch to natively support variable image sizes and aspect ratios. Block-diagonal attention masks enable sequence packing for batching, while RoPE-2D encodings facilitate variable image sizes. Note that the attention mask and position encodings are fed to the vision transformer as additional input, and utilized only in the self-attention layers.

2 Architectural details

Pixtral 12B is based on the transformer architecture [22], and consists of a *multimodal decoder* to perform high-level reasoning, and a *vision encoder* to allow the model to ingest images. The main parameters of the model are summarized in Table 1.

Parameters	Decoder	Encoder
dim	5120	1024
n_layers	40	24
head_dim	128	64
hidden_dim	14336	4096
n_heads	32	16
n_kv_heads	8	16
context_len	131072	4096
vocab_size	131072	-
patch_size	-	16

Table 1: Decoder and encoder parameters.

2.1 Multimodal Decoder

Pixtral 12B is built on top of Mistral Nemo 12B [15], a 12-billion parameter decoder-only language model that achieves strong performance across a range of knowledge and reasoning tasks.

2.2 Vision Encoder

In order for Pixtral 12B to ingest images, we train a new vision encoder from scratch, named Pixtral-ViT. Here, our goal is to instantiate a simple architecture which is capable of processing images across a wide range of resolutions and aspect ratios. To do this, we build a 400 million parameter vision transformer [5] (see Table 1) and make four key changes over the standard architectures [17]:

Break tokens: In order to assist the model in distinguishing between images with the same number of patches (same area) but different aspect ratios, we include `[IMAGE BREAK]` tokens between image rows [2]. We further include an `[IMAGE END]` token at the end of an image sequence.

Gating in FFN: Instead of standard feedforward layer in the attention block, we use gating in the hidden layer [19].

Sequence packing: In order to efficiently process images within a single batch, we flatten the images along the sequence dimension and concatenate them [3]. We construct a block-diagonal mask to ensure no attention leakage between patches from different images.

RoPE-2D: We replace traditional *learned* and *absolute* position embeddings for image patches with *relative*, *rotary* position encodings [11, 20] in the self-attention layers. While learned position embeddings must be interpolated to deal with new image sizes (often at the cost of performance), relative position encodings lend themselves naturally to variable image sizes.