DeepSeek-OCR 2: Visual Causal Flow

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

Abstract

We present DeepSeek-OCR 2 to investigate the feasibility of a novel encoder—DeepEncoder V2—capable of dynamically reordering visual tokens upon image semantics. Conventional vision-language models (VLMs) invariably process visual tokens in a rigid raster-scan order (top-left to bottom-right) with fixed positional encoding when fed into LLMs. However, this contradicts human visual perception, which follows flexible yet semantically coherent scanning patterns driven by inherent logical structures. Particularly for images with complex layouts, human vision exhibits causally-informed sequential processing. Inspired by this cognitive mechanism, DeepEncoder V2 is designed to endow the encoder with causal reasoning capabilities, enabling it to intelligently reorder visual tokens prior to LLM-based content interpretation. This work explores a novel paradigm: whether 2D image understanding can be effectively achieved through two-cascaded 1D causal reasoning structures, thereby offering a new architectural approach with the potential to achieve genuine 2D reasoning. Codes and model weights are publicly accessible at http://github.com/deepseek-ai/DeepSeek-OCR-2.

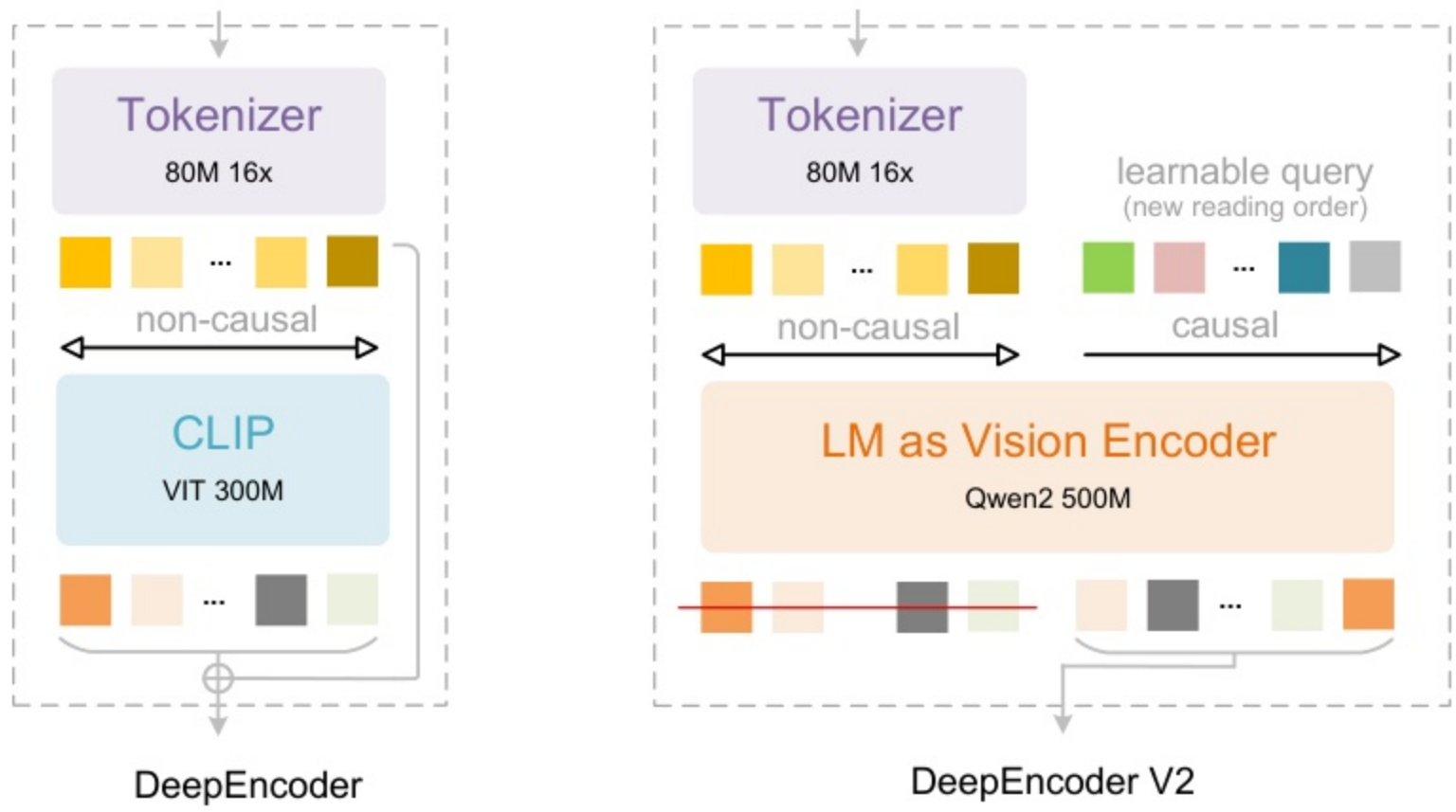

Figure 1 | We substitute the CLIP component in DeepEncoder with an LLM-style architecture. By customizing the attention mask, visual tokens utilize bidirectional attention while learnable queries adopt causal attention. Each query token can thus attend to all visual tokens and preceding queries, allowing progressive causal reordering over visual information.

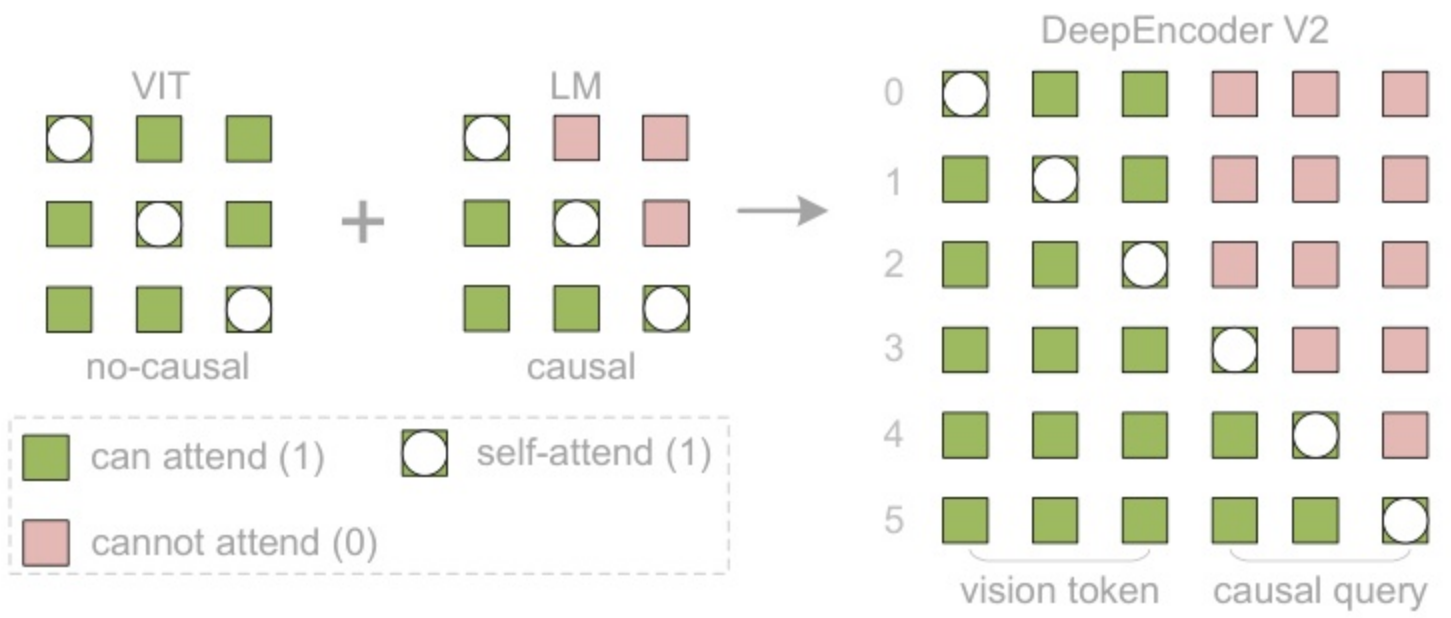

Figure 5 | Attention mask architecture of DeepEncoder V2. Concatenation of bidirectional mask (vision tokens, ViT-like) and causal triangular mask (flow tokens, LLM decoder-style).

number of crops k ranging from 0 to 6 (no cropping is applied when both image dimensions are smaller than 768). All local views share a unified set of 144 query embeddings, denoted as $\text{query}_{\text{local}}$. Therefore, the total number of reordered visual tokens fed to the LLM is $k \times 144 + 256$, ranging from $[256, 1120]$. This maximum token count (1120) is lower than DeepSeek-OCR's 1156 (Gundam mode) and matches Gemini-3-Pro's maximum visual token budget.

3.2.4. Attention mask

To better illustrate the attention mechanism of DeepEncoder V2, we visualize the attention mask in Figure 5. The attention mask is composed of two distinct regions. The left region applies bidirectional attention (similar to ViT) to original visual tokens, allowing full token-to-token visibility. The right region employs causal attention (triangular mask, identical to decoder-only LLMs) for causal flow tokens, where each token attends only to previous tokens. These two components are concatenated along the sequence dimension to construct DeepEncoder V2's attention mask (M), as follows:

$$M = \begin{bmatrix} \mathbf{1}_{m \times m} & \mathbf{0}_{m \times n} \\ \mathbf{1}_{n \times m} & \text{LowerTri}(n) \end{bmatrix}, \quad \text{where } n = m \tag{1}$$

where n is the number of causal query tokens, m represents vanilla visual tokens number, and LowerTri denotes a lower triangular matrix (with ones on and below the diagonal, zeros above).

3.3. DeepSeek-MoE Decoder

Since DeepSeek-OCR 2 primarily focuses on encoder improvements, we do not upgrade the decoder component. Following this design principle, we retain DeepSeek-OCR's decoder – a 3B-parameter MoE structure with about 500M active parameters. The core forward pass of DeepSeek-OCR 2 can be formulated as:

$$\mathbf{O} = \mathcal{D}\left(\pi_Q\left(\mathcal{T}^L\left(\mathcal{E}(\mathbf{I}) \oplus \mathbf{Q}_0; \mathbf{M}\right)\right)\right) \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ is the input image, $\mathcal{E}$ is the vision tokenizer mapping images to m visual tokens $\mathbf{V} \in \mathbb{R}^{m \times d}$, $\mathbf{Q}_0 \in \mathbb{R}^{n \times d}$ are learnable causal query embeddings, $\oplus$ denotes sequence concatenation, $\mathcal{T}^L$ represents an L-layer Transformer with masked attention, $\mathbf{M} \in \{0, 1\}^{2n \times 2n}$ is the block causal attention mask defined in Equation 1, π_Q is the projection operator that extracts the last n tokens (i.e., $\mathbf{Z} = \mathbf{X}_{m+1:m+n}$), $\mathcal{D}$ is the language decoder, and $\mathbf{O} \in \mathbb{R}^{n \times |\mathcal{V}|}$ is the output logits over LLM vocabulary.