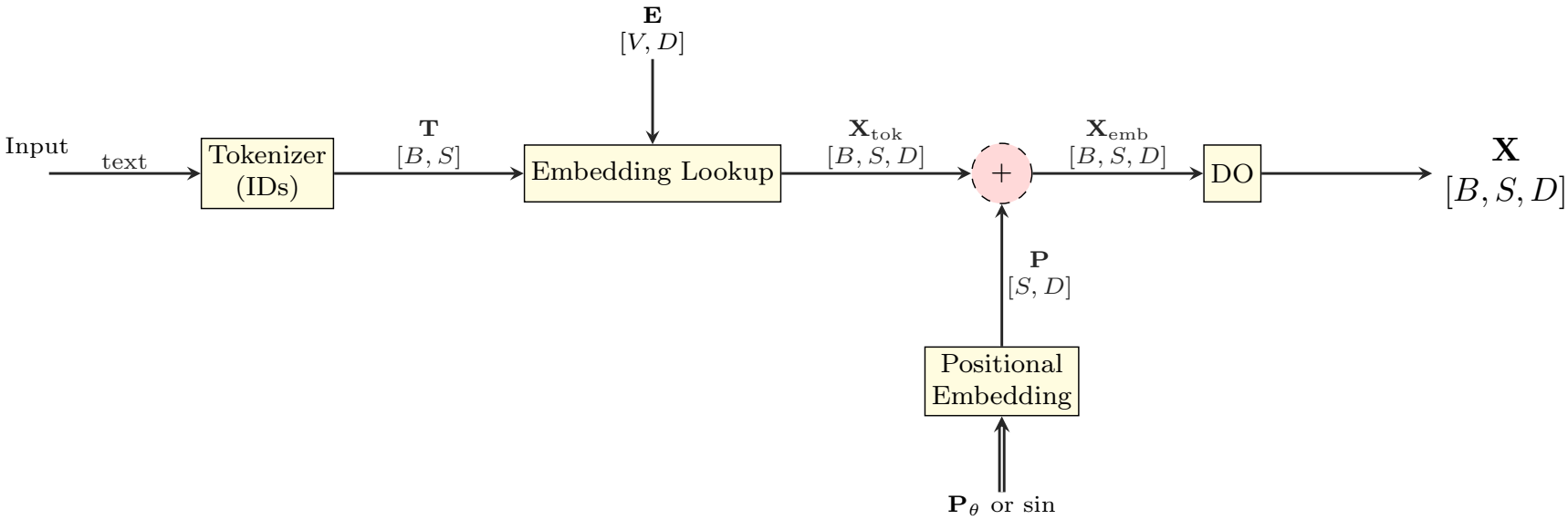
Input → Embedding → LN (Input to MHA)



Operations (Ops)

Abbrev	Name	Type / Shape	Notes
Tokenizer	Tokenizer (IDs)	op	Maps raw text $\rightarrow$ integer ids $\mathbf{T} \in \mathbb{Z}^{[B,S]}$.
Embedding Lookup	Embedding Lookup	op	Gathers rows from $\mathbf{E} \in \mathbb{R}^{V \times D}$ using ids $\mathbf{T}$.
$+$	Element-wise Add (dashed circle)	op	Adds token and positional embeddings; broadcasting over B, S if needed.
DO	Dropout	op	Training-time stochastic dropout on $\mathbf{X}_{\text{emb}}$; identity at inference.
(none)	Broadcast $\text{BC}_{B,S}(\cdot)$	op	Expands $[S, D]$ (or $[D]$) to $[B, S, D]$ across batch/sequence.

Data Tensors (Values)

Symbol	Name	Shape	Notes
text	Raw input text	—	Character/byte stream before tokenization.
$\mathbf{T}$	Token ids	$[B, S]$	Output of Tokenizer; integers in $\{0, \ldots, V-1\}$.
$\mathbf{E}$	Embedding matrix (params)	$[V, D]$	Trainable; each vocab entry has a D-dim vector.
$\mathbf{X}_{\text{tok}}$	Token embeddings	$[B, S, D]$	lookup$(\mathbf{E}, \mathbf{T})$.
$\mathbf{P}$	Positional embedding	$[S, D]$ (or $[B, S, D]$)	Learned $\mathbf{P}_\theta$ **or** sinusoidal (fixed); broadcast to $[B, S, D]$.
$\mathbf{X}_{\text{emb}}$	Sum of token+pos	$[B, S, D]$	$\mathbf{X}_{\text{tok}} + \text{BC}_{B,S}(\mathbf{P})$.
$\mathbf{X}$	Input to MHA	$[B, S, D]$	After dropout (DO); goes to LN/MHA stack.
$\mathbf{P}_\theta$	Learned pos. params	matches $\mathbf{P}$	Used when positions are trainable; otherwise "sin" denotes fixed sinusoidal.

Shape symbols: B=batch size, S=sequence length, D=model dim, V=vocab size.

Notes: In practice, $\mathbf{P}$ may be pre-broadcast to $[B, S, D]$ or added per-token with implicit broadcasting.