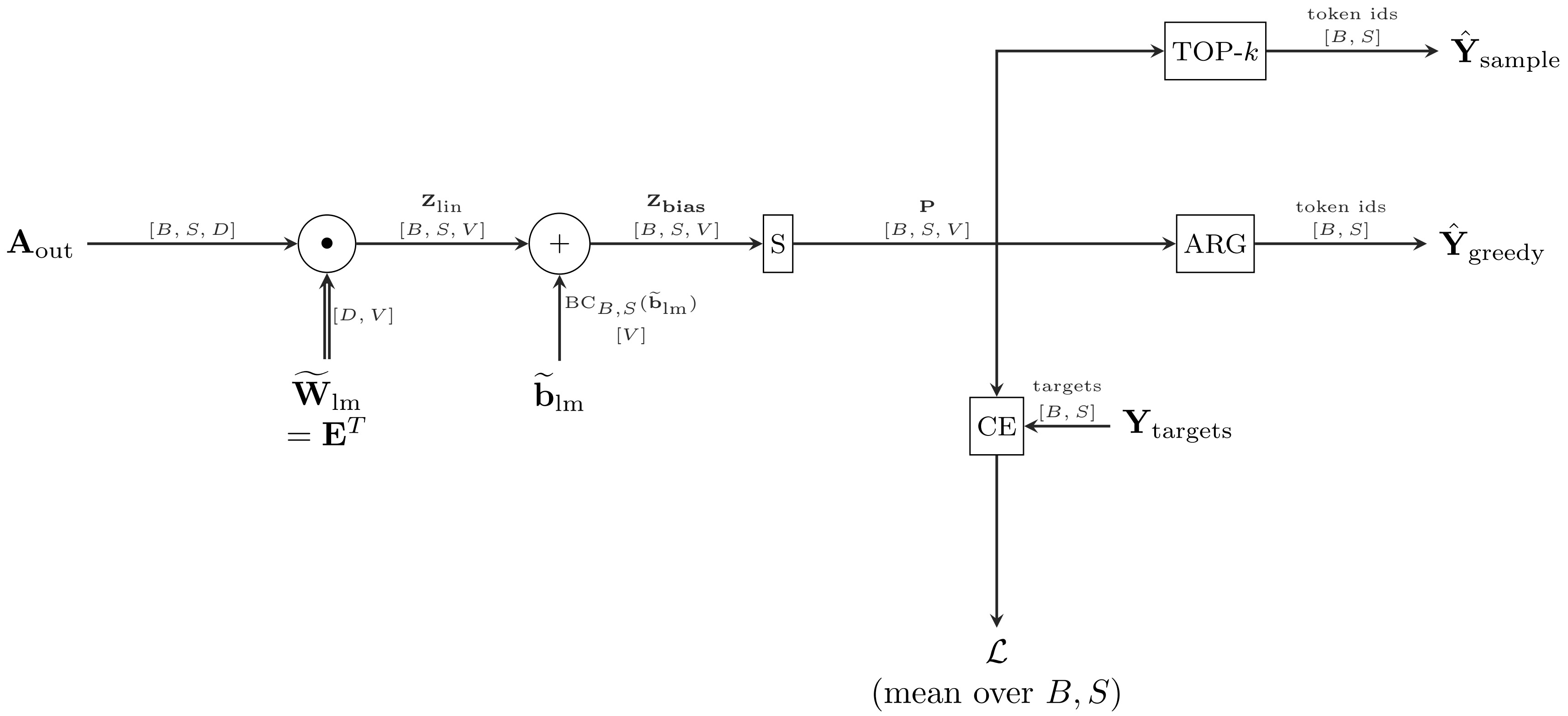
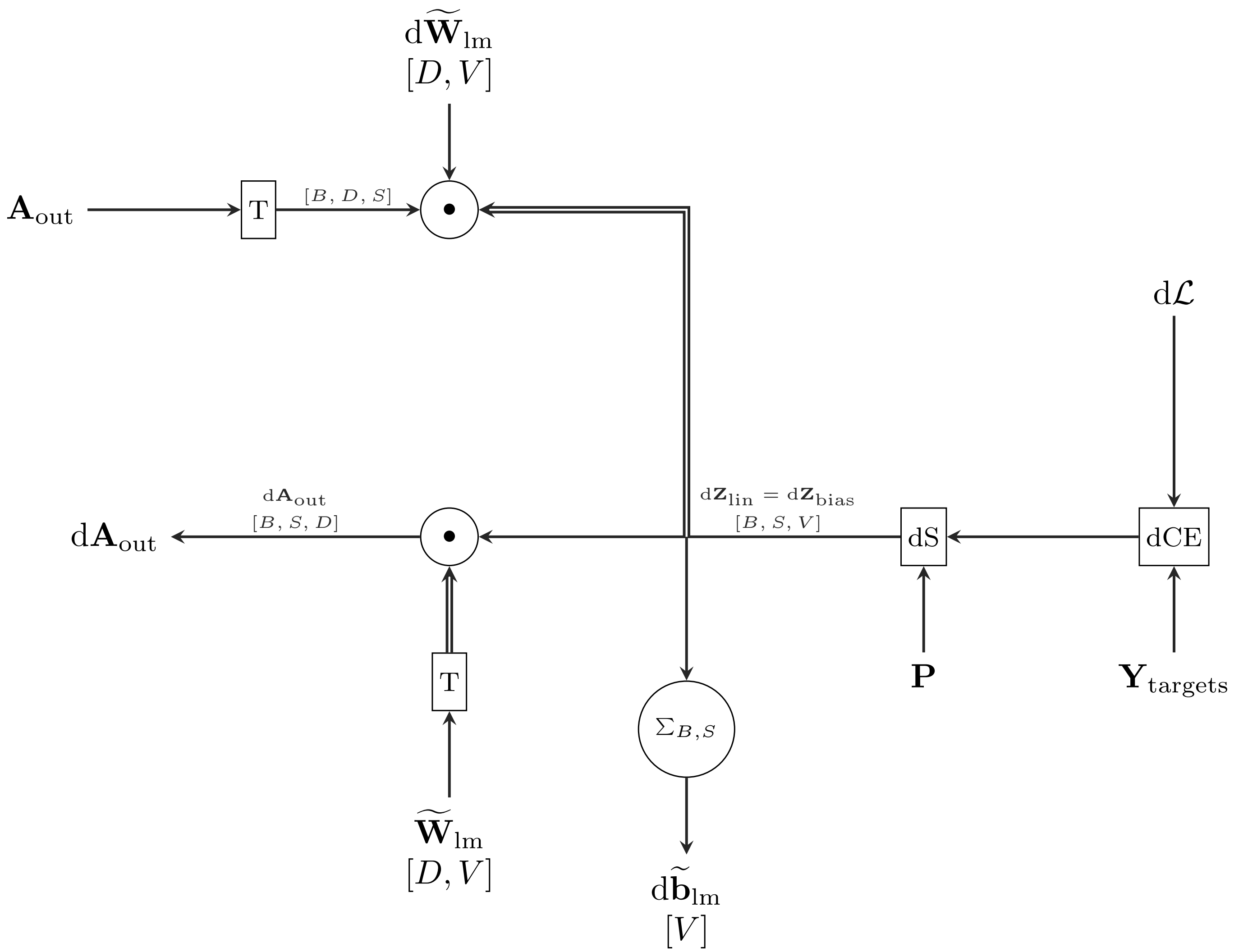
Token Generation & Loss (Forward)


$\mathbf{A}_{\text{out}}$
$[B, S, D]$
$\odot$
$[D, V]$
$\widetilde{\mathbf{W}}_{\text{lm}}$
$= \mathbf{E}^T$
$\mathbf{z}_{\text{lin}}$
$[B, S, V]$
$+$
$\text{BC}_{B,S}(\widetilde{\mathbf{b}}_{\text{lm}})$
$[V]$
$\widetilde{\mathbf{b}}_{\text{lm}}$
$\mathbf{z}_{\textbf{bias}}$
$[B, S, V]$
S
$\mathbf{P}$
$[B, S, V]$
TOP-k
token ids
$[B, S]$
$\hat{\mathbf{Y}}_{\text{sample}}$
ARG
token ids
$[B, S]$
$\hat{\mathbf{Y}}_{\text{greedy}}$
CE
targets
$[B, S]$
$\mathbf{Y}_{\text{targets}}$
$\mathcal{L}$
(mean over B, S)

Token Generation & Loss — Backward (Corrected)


$\mathrm{d}\widetilde{\mathbf{W}}_{\mathrm{lm}}$
$[D, V]$
$\mathbf{A}_{\mathrm{out}}$
T
$[B, D, S]$
$\odot$
$\mathrm{d}\mathcal{L}$
$\mathrm{d}\mathbf{A}_{\mathrm{out}}$
$[B, S, D]$
$\mathrm{d}\mathbf{A}_{\mathrm{out}}$
$\odot$
$\mathrm{d}\mathbf{Z}_{\mathrm{lin}} = \mathrm{d}\mathbf{Z}_{\mathrm{bias}}$
$[B, S, V]$
dS
dCE
T
$\Sigma_{B,S}$
$\mathbf{P}$
$\mathbf{Y}_{\mathrm{targets}}$
$\widetilde{\mathbf{W}}_{\mathrm{lm}}$
$[D, V]$
$\mathrm{d}\widetilde{\mathbf{b}}_{\mathrm{lm}}$
$[V]$

Operations (Ops)

Abbrev	Name	Type / Shape	Notes
S	Softmax	op	Over vocab axis V; outputs probabilities $\mathbf{P}$.
CE	Cross-Entropy	op	Usually *sparse* CE consuming label indices $\mathbf{Y}$.
ARG	Argmax (greedy)	op	argmax_V to get token ids (no gradient).
TOP-k	Top-k / sampling	op	Optional decoding path (or nucleus sampling); no gradient.
T	Transpose	op	E.g., $\widetilde{\mathbf{W}}_{\text{lm}}^T \in \mathbb{R}^{V \times D}$.
$\text{BC}_{B,S}(\cdot)$	Broadcast	op	Expand $[V] \to [B, S, V]$ for bias add.
dS	Softmax backward	op	The output is $\mathrm{d}\mathbf{Z}_{\text{bias}} = \mathbf{P} - \text{onehot}(\mathbf{Y})$ (with CE).
dAddB	Addition (Bias) backward	op	Passes $\mathrm{d}\mathbf{Z}_{\text{bias}}$ to $\mathrm{d}\mathbf{Z}_{\text{lin}}$ and $\sum_{B,S}$.
$\sum_{B,S}$	Summation	op	Sums $\mathrm{d}\mathbf{Z}_{\text{bias}}$ over axes B and S to get $\mathrm{d}\widetilde{\mathbf{b}}_{\text{lm}}$.

Data Tensors (Values)

Symbol	Name	Shape	Notes
$\mathbf{A}_{\text{out}}$	Transformer output (hidden)	$[B, S, D]$	Final hidden from the Transformer block(s).
$\widetilde{\mathbf{W}}_{\text{lm}}$	LM head weight (tied)	$[D, V]$	Typically tied to $\mathbf{E}^T$.
$\widetilde{\mathbf{b}}_{\text{lm}}$	LM head bias	$[V]$	Broadcast-added over $[B, S, V]$.
$\mathbf{Z}_{\text{lin}}$	Logits (linear output)	$[B, S, V]$	$\mathbf{Z}_{\text{lin}} = \mathbf{A}_{\text{out}} \widetilde{\mathbf{W}}_{\text{lm}}$.
$\mathbf{Z}_{\text{bias}}$	Logits (final/Softmax input)	$[B, S, V]$	$\mathbf{Z}_{\text{bias}} = \mathbf{Z}_{\text{lin}} + \widetilde{\mathbf{b}}_{\text{lm}}$.
$\mathbf{P}$	Probabilities	$[B, S, V]$	$\mathbf{P} = \text{softmax}(\mathbf{Z}_{\text{bias}})$.
$\mathbf{Y}$	Target token ids	$[B, S]$	Ground-truth indices (sparse labels).
$\mathcal{L}$	Loss	scalar or $[B, S]$	Typically mean over B, S.
$\mathrm{d}\mathcal{L}$	Loss gradient	scalar-grad	Starting signal for backward pass.
$\mathrm{d}\mathbf{Z}_{\text{bias}}$	Final Logits gradient	$[B, S, V]$	From CE+Softmax: $\mathbf{P} - \text{onehot}(\mathbf{Y})$.
$\mathrm{d}\mathbf{Z}_{\text{lin}}$	Linear output grad	$[B, S, V]$	Same as $\mathrm{d}\mathbf{Z}_{\text{bias}}$ (input to $\mathbf{Z}_{\text{lin}}$ matmul).
$\mathrm{d}\widetilde{\mathbf{W}}_{\text{lm}}$	LM weight grad	$[D, V]$	$= \mathbf{A}_{\text{out}}^T \mathrm{d}\mathbf{Z}_{\text{lin}}$.
$\mathrm{d}\widetilde{\mathbf{b}}_{\text{lm}}$	LM bias grad	$[V]$	$= \sum_{B,S}(\mathrm{d}\mathbf{Z}_{\text{bias}})$.
$\mathrm{d}\mathbf{A}_{\text{out}}$	Hidden grad	$[B, S, D]$	$= \mathrm{d}\mathbf{Z}_{\text{lin}} \widetilde{\mathbf{W}}_{\text{lm}}^T$.

Shapes: B=batch, S=sequence length, D=hidden dim, V=vocab size.