Based on the DynamicViT paper, the loss function that constrains the pruning ratio is implemented as an MSE loss between the target pruning ratio and the actual pruning ratio at each stage. The mathematical implementation is:

L_ratio = (1/BS) * :5ö#Ó Fò " £_s=1 to S (< â‡2’ Ò ƒ ôâ“£_i=1 to N D_b,s_i)^2

Where:
- B is batch size
- S is number of pruning stages (typically 3)
- < â‡2’ —2 F &vWB atio at stage s (e.g., [< Â Á^2, < ã5Ò•
- N is number of tokens
- D_b,s_i is binary decision mask for sample b at stage s
- The term (1/N):5ö“Ó Fò â Eö ,s_i calculates actual keeping ratio

This loss is combined with other losses (classification, distillation, KL divergence) with a weighting factor ;µ÷ atio (set to 2 in their experiments).

The full training objective is:
L = L_cls + ;µô´Â¤Åô´Â ² »_distill*L_distill + ;µ÷ atio*L_ratio

This MSE loss ensures the model maintains the desired pruning ratio at each stage while allowing some flexibility for input-dependent pruning decisions.