Based on the paper "DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification", the loss function that constrains the pruning proportion is implemented as a Mean Squared Error (MSE) loss between the target pruning ratio and the actual pruning ratio at each stage. Here's the mathematical implementation:

The ratio constraint loss (L_ratio) is defined as:

L_ratio = (1/BS) * :5ö#Ó ä" £_s=1^S (< â‡2’ Ò ƒ ôâ’ ¢ £_i=1^N D0%ö•â† ,s))^2

Where:
- B is batch size
- S is number of sparsification stages
- < â‡2’ —2 F†R F &vWB °eeping ratio at stage s
- N is number of tokens
- D0%ö•â† ,s) is the binary decision mask (0 or 1) for token i in sample b at stage s

This loss ensures the model maintains the desired pruning ratio <   B V 6, 7F vR v†–ÆR allowing dynamic adjustment based on input content. The full loss combines this with classification loss (L_cls), distillation loss (L_distill), and KL divergence loss (L_KL):

L_total = L_cls + ;µô´Â¤Åô´Â ² »_distill*L_distill + ;µ÷ atio*L_ratio

Where they used ;µô´Â Ò ãRÂ »_distill = 0.5, ;µ÷ atio = 2 in their experiments.

The implementation uses Gumbel-Softmax to make the discrete pruning decisions differentiable during training, while during inference they simply keep the top-k tokens based on the predicted keeping probabilities.