The paper "DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification" introduces a token ratio constraint loss (Equation 15 in the paper) to control the proportion of pruned tokens during training. Here's how it's mathematically modeled:

1. **Mathematical Formulation**:
The token ratio constraint loss is implemented as an MSE loss between the target pruning ratio and the actual proportion of kept tokens at each sparsification stage:

$$
L_{ratio} = \frac{1}{BS} \sum_{b=1}^{B} \sum_{s=1}^{S} \left(\rho^{(s)} - \frac{1}{N} \sum_{i=1}^{N} \hat{D}_{i}^{b,s} \right)^2
$$

Where:
- B = batch size
- S = number of sparsification stages
- $\rho^{(s)}$ = target keeping ratio at stage s
- $\hat{D}^{b,s}$ = binary decision mask for sample b at stage s
- N = total number of tokens

2. **Dynamic Threshold Adjustment**:
The paper uses:
- Fixed predefined target ratios during training (set as a geometric sequence [< Â Á², < 5Ò•
- Dynamic token selection during inference based on predicted importance scores

Key implementation details:
- The threshold isn't dynamically adjusted during training - the target ratios are fixed hyperparameters
- During inference, they select the top-k tokens (k = < â’ & 6VB öâ &VF–7FVB –× ÷ tance scores
- This creates an adaptive threshold that varies per input instance while maintaining the desired global sparsity ratio

The combination of fixed ratio targets during training and dynamic selection during inference allows the model to learn input-adaptive pruning while maintaining control over computational budgets.