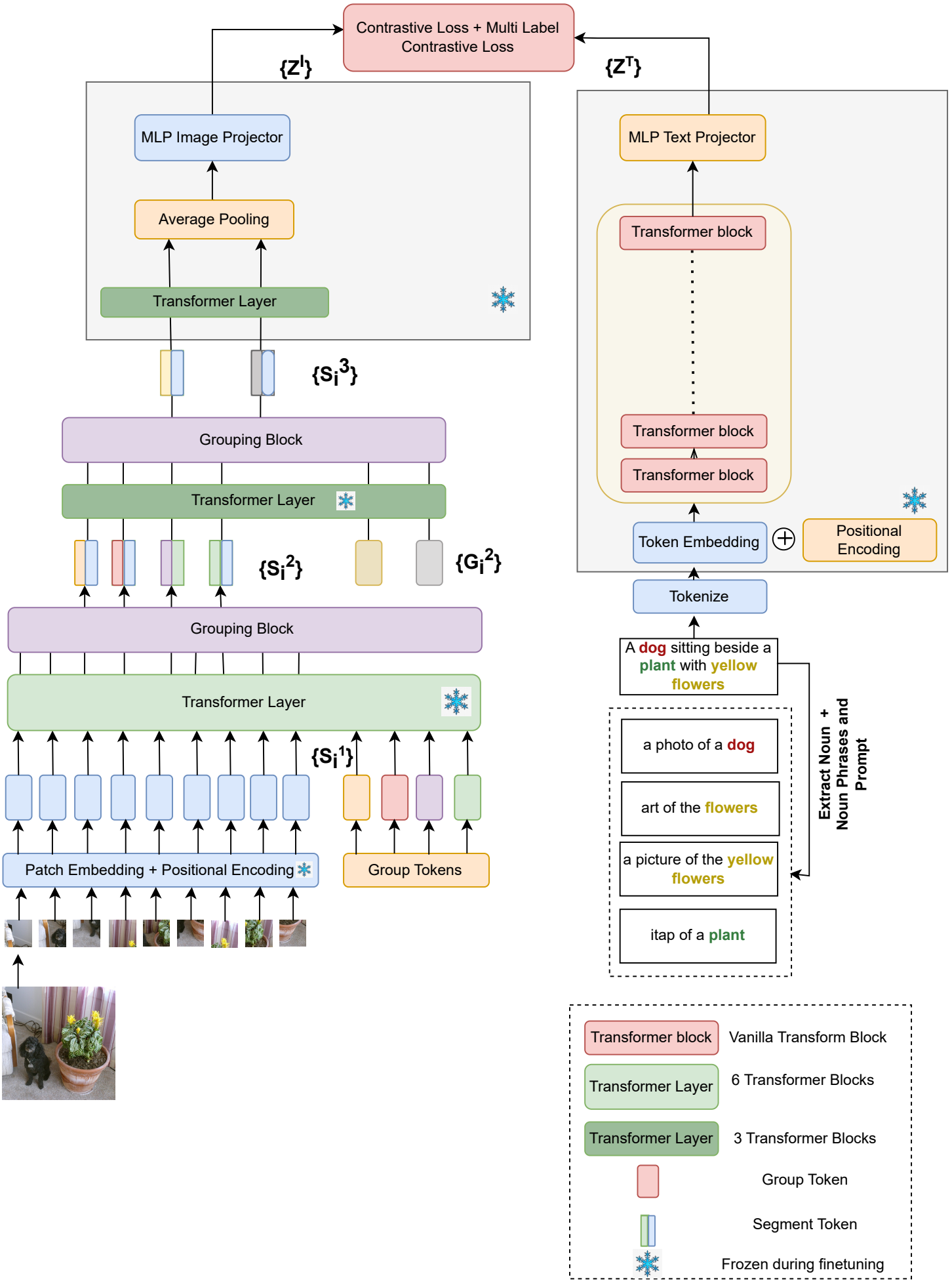
Contrastive Loss + Multi Label Contrastive Loss
{Z^I}
{Z^T}
MLP Image Projector
Average Pooling
Transformer Layer
{S_i^3}
Grouping Block
Transformer Layer
{S_i^2}
{G_i^2}
Grouping Block
Transformer Layer
{S_i^1}
Patch Embedding + Positional Encoding
Group Tokens
MLP Text Projector
Transformer block
Transformer block
Transformer block
Token Embedding
Positional Encoding
Tokenize
A dog sitting beside a plant with yellow flowers
a photo of a dog
art of the flowers
a picture of the yellow flowers
itap of a plant
Extract Noun + Noun Phrases and Prompt
Transformer block
Vanilla Transform Block
Transformer Layer
6 Transformer Blocks
Transformer Layer
3 Transformer Blocks
Group Token
Segment Token
Frozen during finetuning