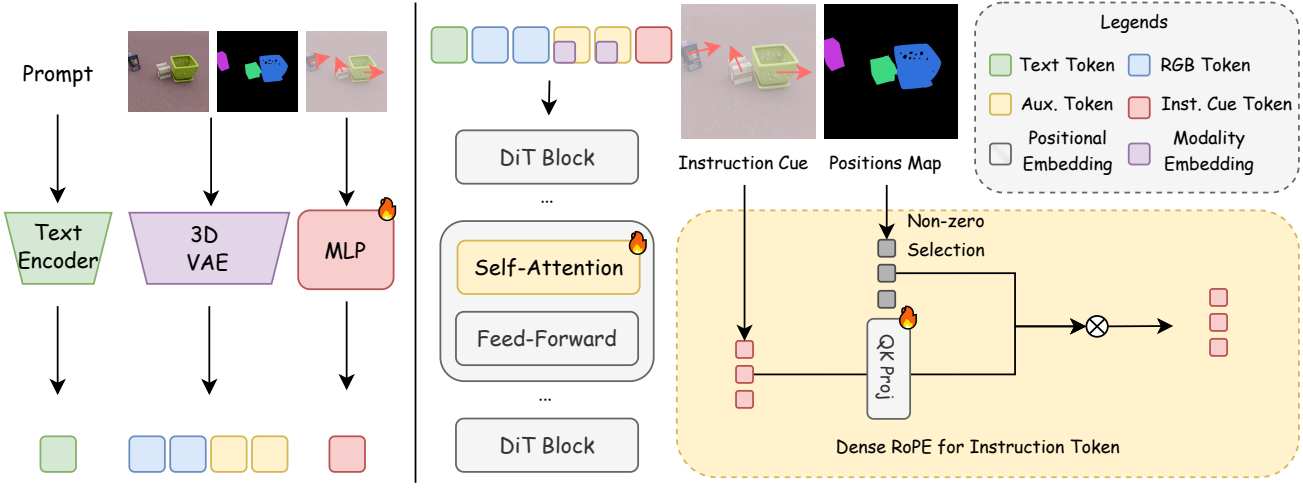
Prompt
Text Encoder
3D VAE
MLP
DiT Block
...
Self-Attention
Feed-Forward
...
DiT Block
Instruction Cue
Positions Map
Legends
Text Token
RGB Token
Aux. Token
Inst. Cue Token
Positional Embedding
Modality Embedding
Non-zero Selection
QK Proj
Dense RoPE for Instruction Token