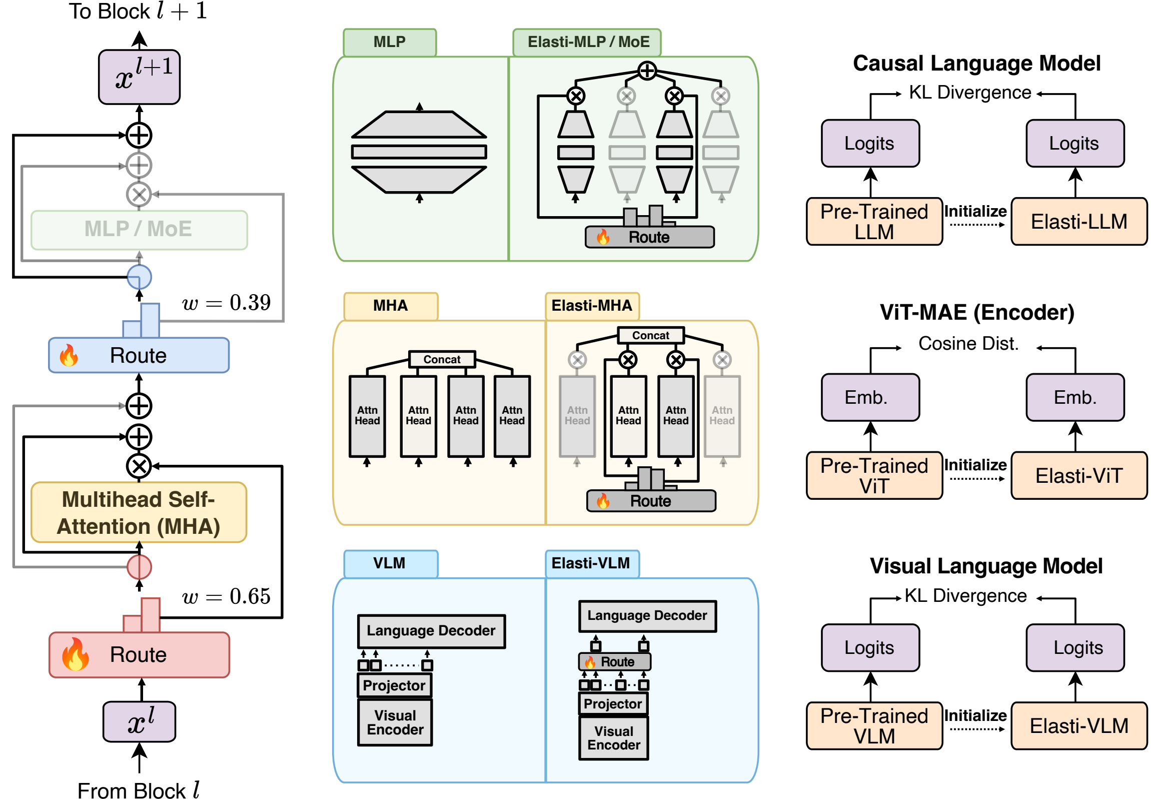
To Block $l+1$
x^{l+1}
MLP / MoE
$w = 0.39$
Route
Multihead Self-Attention (MHA)
$w = 0.65$
Route
x^l
From Block l
MLP
Elasti-MLP / MoE
Route
MHA
Elasti-MHA
Concat
Attn Head
Attn Head
Attn Head
Attn Head
Concat
Attn Head
Attn Head
Attn Head
Attn Head
Route
VLM
Elasti-VLM
Language Decoder
Projector
Visual Encoder
Language Decoder
Route
Projector
Visual Encoder
Causal Language Model
KL Divergence
Logits
Logits
Pre-Trained LLM
Initialize
Elasti-LLM
ViT-MAE (Encoder)
Cosine Dist.
Emb.
Emb.
Pre-Trained ViT
Initialize
Elasti-ViT
Visual Language Model
KL Divergence
Logits
Logits
Pre-Trained VLM
Initialize
Elasti-VLM