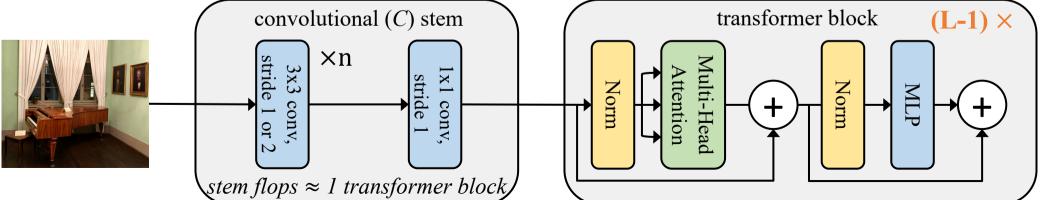


Original ViT (baseline, termed ViT_P): Sensitive to lr and wd choice

- Converges slowly
- Works with AdamW, but not SGD
- **Underperforms** sota CNNs on ImageNet



Ours (termed ViT_C, same runtime):

- ✓ *Robust to Ir and wd choice*
- ✓ Converges quickly
- Works with AdamW, and also SGD
- ✓ Outperforms sota CNNs on ImageNet