Method	Language		Action		Vision	Quantization	Training-Free
	Attention	MLP	Attention	MLP			
TinyVLA	✓ Compact Transformer	✗ Compact Transformer	✓ Compact DP	✓ Compact DP	✗	✗	✗
EfficientVLA	✓ Layer drop/pruning	✓ Layer pruning	✓ Feature cache	✗ Feature cache	✓ Token reuse + feature cache	✗	✓
VLA-Cache	✗	✗	✗	✗	✓ KV reuse for static vision tokens	✗	✓
MoLe-VLA	✓ Dynamic routing	✓ Dynamic routing	✗	✗	✗	✗	✗
QuantVLA	✓ Full linear Attn	✓ Full linear MLP	✓ Full linear Attn	✓ Full linear MLP	✗ Keep vision frozen	✓ PTQ	✓