ONE UNIFIED-MEMORY BUDGET

Target + resident draft | KV + recurrent state | Working memory | Visual / host state

01 KVExec
Compact pages → bounded operands
Rotated attention · coalesced staging

02 PhaseSwap
Last-consumer ownership
Workspace reuse · shared verify rows

03 StateTrans
Committed KV + recurrent + draft state
Schedule · publish · persist · restore

04 CrossFire
CPU/GPU partition · GPU/ANE overlap
Phase-aware admission and fallback