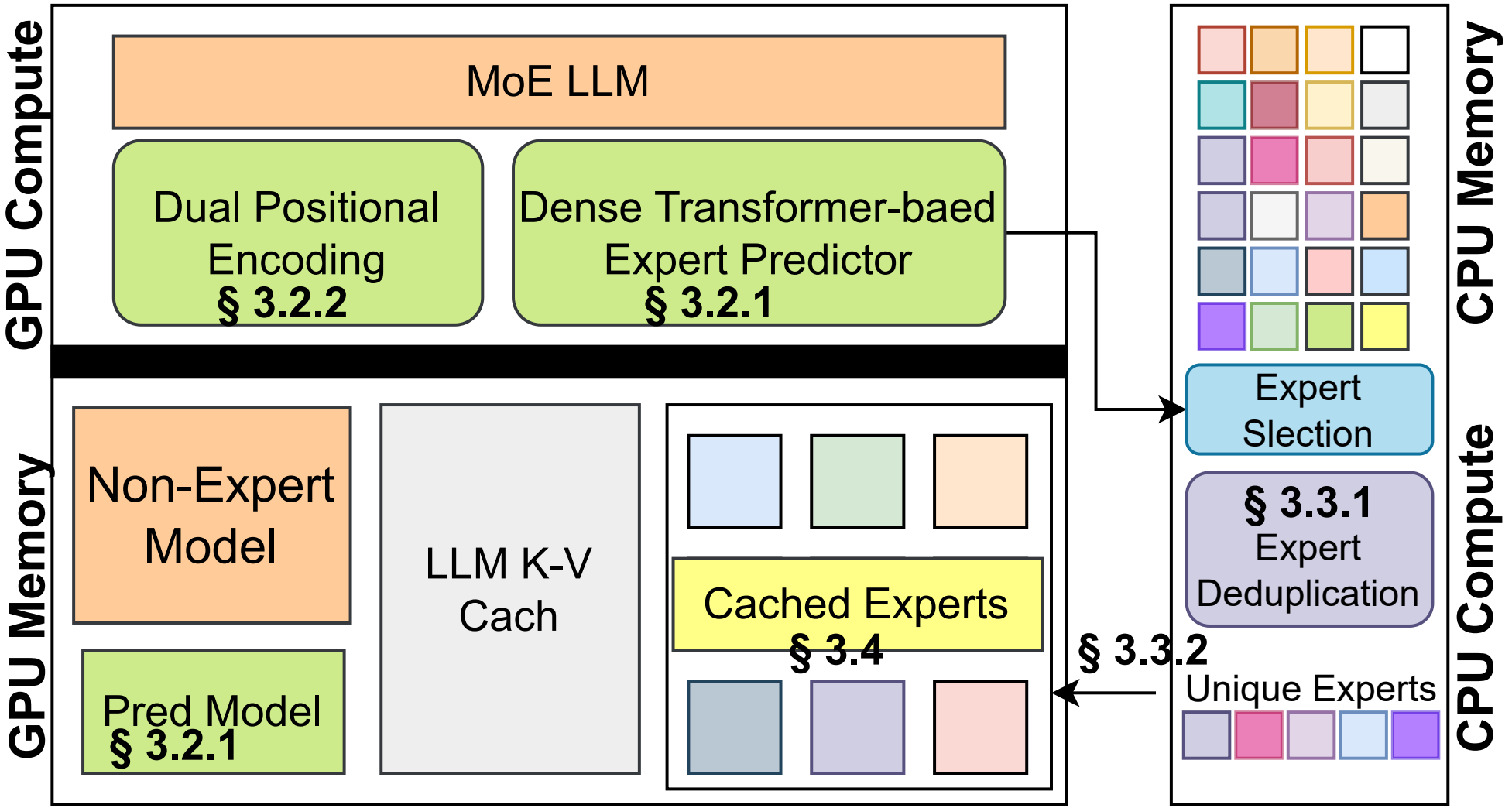
GPU Compute
MoE LLM
Dual Positional Encoding
§ 3.2.2
Dense Transformer-baed Expert Predictor
§ 3.2.1
GPU Memory
Non-Expert Model
Pred Model
§ 3.2.1
LLM K-V Cach
Cached Experts
§ 3.4
CPU Memory
Expert Slection
§ 3.3.1
Expert Deduplication
§ 3.3.2
Unique Experts
CPU Compute