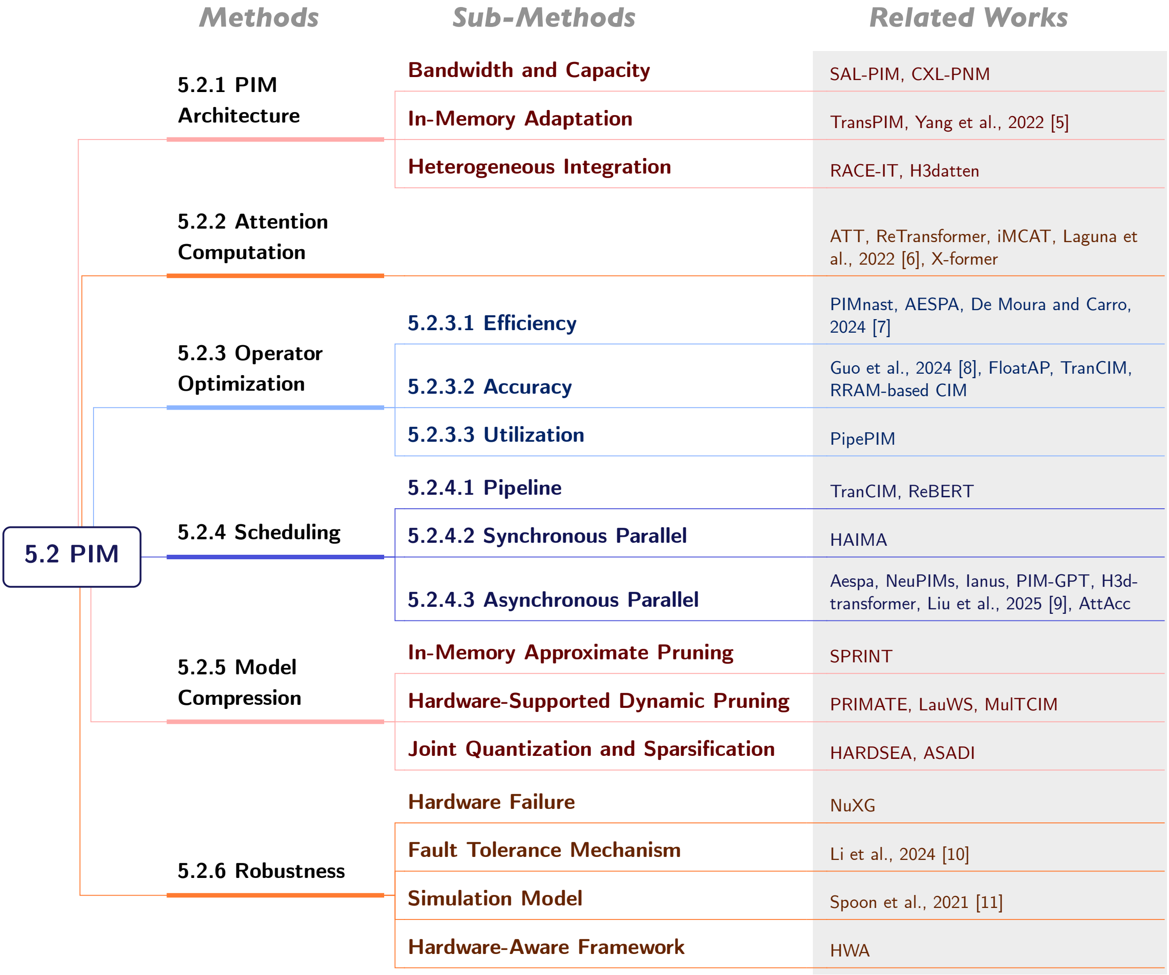

Methods	Sub-Methods	Related Works
5.2 PIM		
5.2.1 PIM Architecture	Bandwidth and Capacity	SAL-PIM, CXL-PNM
	In-Memory Adaptation	TransPIM, Yang et al., 2022 [5]
	Heterogeneous Integration	RACE-IT, H3datten
5.2.2 Attention Computation		ATT, ReTransformer, iMCAT, Laguna et al., 2022 [6], X-former
5.2.3 Operator Optimization	5.2.3.1 Efficiency	PIMnast, AESPA, De Moura and Carro, 2024 [7]
	5.2.3.2 Accuracy	Guo et al., 2024 [8], FloatAP, TranCIM, RRAM-based CIM
	5.2.3.3 Utilization	PipePIM
5.2.4 Scheduling	5.2.4.1 Pipeline	TranCIM, ReBERT
	5.2.4.2 Synchronous Parallel	HAIMA
	5.2.4.3 Asynchronous Parallel	Aespa, NeuPIMs, Ianus, PIM-GPT, H3d-transformer, Liu et al., 2025 [9], AttAcc
5.2.5 Model Compression	In-Memory Approximate Pruning	SPRINT
	Hardware-Supported Dynamic Pruning	PRIMATE, LauWS, MulTCIM
	Joint Quantization and Sparsification	HARDSEA, ASADI
5.2.6 Robustness	Hardware Failure	NuXG
	Fault Tolerance Mechanism	Li et al., 2024 [10]
	Simulation Model	Spoon et al., 2021 [11]
	Hardware-Aware Framework	HWA