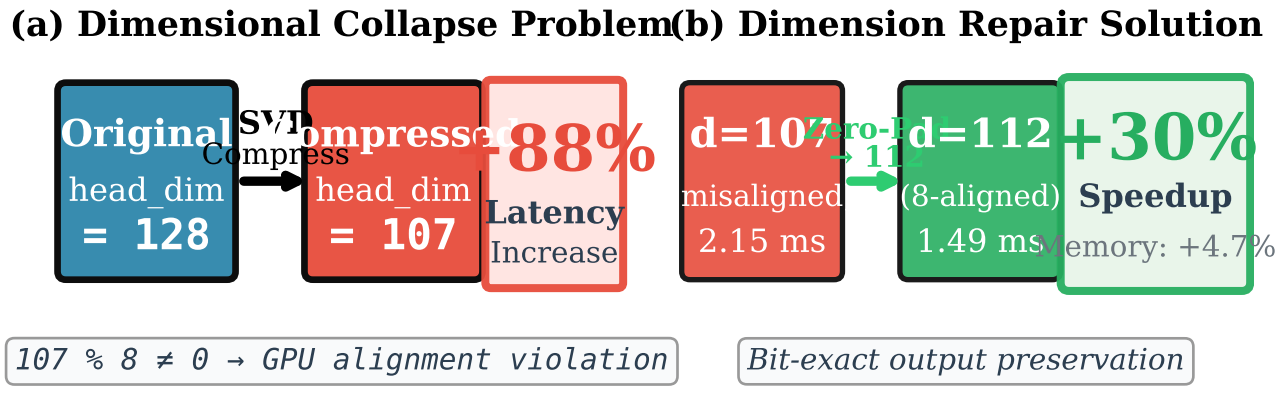
(a) Dimensional Collapse Problem
(b) Dimension Repair Solution
Original
head_dim
= 128
SVD
Compress
Compressed
head_dim
= 107
-88%
Latency
Increase
d=107
misaligned
2.15 ms
Zero-Pad
→ 112
d=112
(8-aligned)
1.49 ms
+30%
Speedup
Memory: +4.7%
107 % 8 ≠ 0 → GPU alignment violation
Bit-exact output preservation