Qwen2.5-VL-3B

Result 0: non fine-tuned qwen2.5-VL

{'bleu': {'bleu': 0.1658110055141822, 'precisions': [0.5483313833639757, 0.40979425755057597, 0.3251519263673373, 0.2698340324639796], 'brevity\_penalty': 0.4425021104673164, 'length\_ratio': 0.5508700856274744, 'translation\_length': 35898, 'reference\_length': 65166}, 'exact\_match': {'exact\_match': np.float64(0.0)}, 'average\_inference\_time(s)': 0.8212410094738006, 'max\_gpu\_memory\_allocated(MB)': 7754.875}

Result 2: 4-bit use double quant, kv cache

{'bleu': {'bleu': 0.34382248219728506, 'precisions': [0.5338621853417532, 0.4217021050581591, 0.3511345817637835, 0.3017271442468293], 'brevity\_penalty': 0.8748901815864152, ength\_ratio': 0.8821010956633828, 'translation\_length': 57483, 'reference\_length': 65166}, 'exact\_match': {'exact\_match': np.float64(0.0)}, 'average\_inference\_time': 2.34772648962402, 'max\_gpu\_memory\_allocated': 3878.0}

Result 3: no quantized, kv cache

{'bleu': {'bleu': 0.8958983669497077, 'precisions': [0.9648813517013277, 0.9300600662472067, 0.8929094803291624, 0.8594389341713313], 'brevity\_penalty': 0.9834604577177702, 'length\_ratio': 0.9835957401098733, 'translation\_length': 64097, 'reference\_length': 65166}, 'exact\_match': {'exact\_match': np.float64(0.376)}, 'average\_inference\_time': 2.315101310014725, 'max\_gpu\_memory\_allocated': 7754.875}

Result 4: 8-bit quantized, kv cache

{'bleu': {'bleu': 0.89536274694, 'precisions': [0.9646537879, 0.928837290336, 0.891449538922, 0.8579443226798], 'brevity\_penalty': 0.984084234544, 'length\_ratio': 0.984209557131, 'translation\_length': 64137, 'reference\_length': 65166}, 'exact\_match': {'exact\_match': np.float64(0.369)}, 'average\_inference\_time(s)': 7.942765971899, 'max\_gpu\_memory\_allocated(MB)': 4025.552}

Result 5: no quantized, kv cache, spda attention

{'bleu': {'bleu': 0.8964192548244294, 'precisions': [0.9661998657361872, 0.9311531568680317, 0.8940421897410278, 0.8605660562780489], 'brevity\_penalty': 0.9827738608454722, 'length\_ratio': 0.9829205413866127, 'translation\_length': 64053, 'reference\_length': 65166}, 'exact\_match': {'exact\_match': np.float64(0.378)}, 'average\_inference\_time(s)': 2.0987275393009184, 'max\_gpu\_memory\_allocated(MB)': 7754.875}