

Results

Table 1: Training Losses

	CE Loss	Alignment Loss
First Stage	2.046	0.514
Second Stage	0.819	—

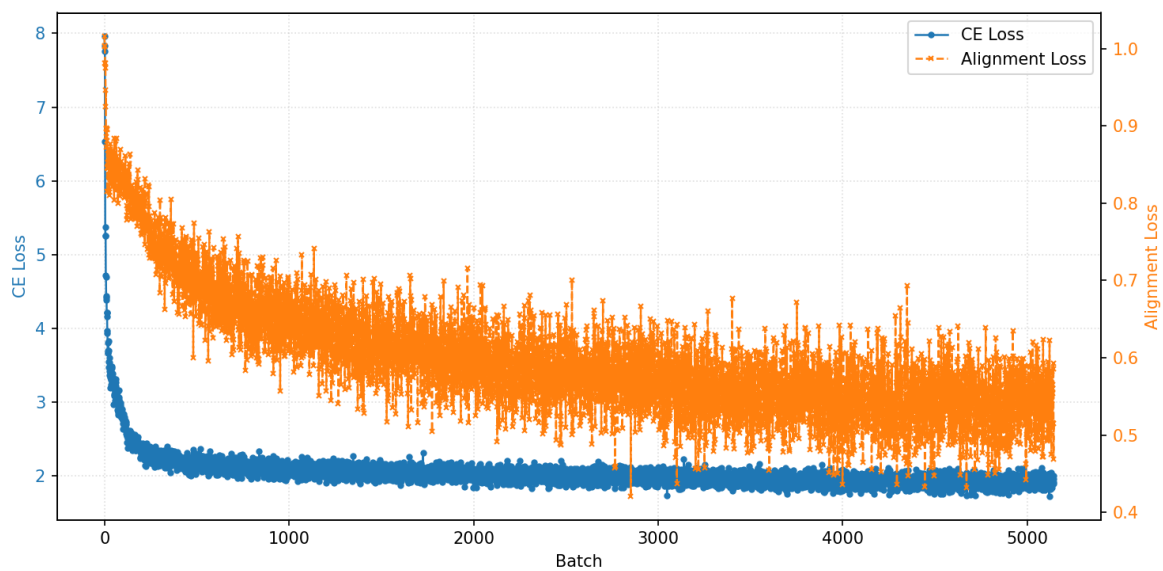
Dataset	Perception Score	Cognition Score	LLaVA Baseline
mme			1510.8/348.2

Dataset	Exact Match	stderr	LLaVA Baseline
aizd			0.548 ↑
scienceqa_img			0.704 ↑
gqa			0.620 ↑
textvqa_val			

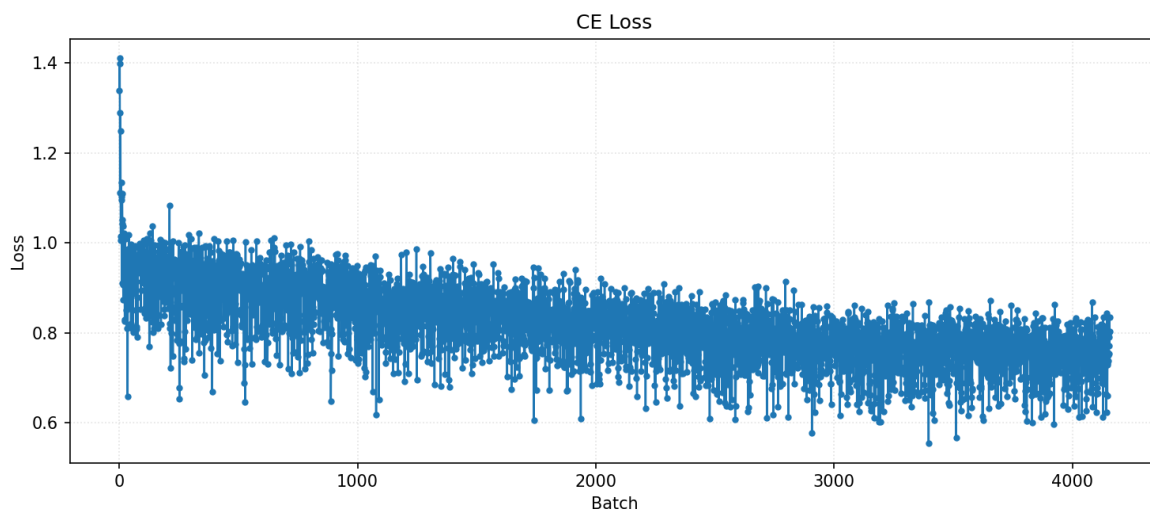
Dataset	Accuracy	LLaVA Baseline
mmmu_val	0.3611 ↑	0.353

Dataset	Accuracy	F1 Score	LLaVA Baseline
pope			0.859 ↑

Dataset	seed_all	image	video	LLaVA Baseline
seedbench				60.5



(a) First Stage Losses



(b) Second Stage Losses