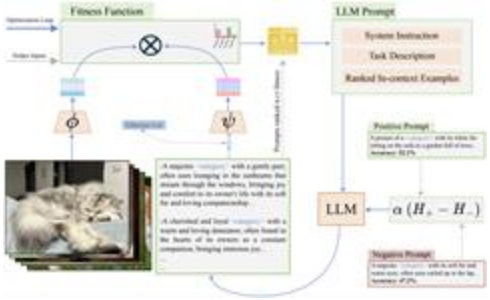
GLOV: Guided Large Language Models as Implicit Optimizers for Vision Language Models

M. Jehanzeb Mirza, James Glass
MIT CSAIL

Abstract

We propose GLOV, a method that leverages Large Language Models (LLMs) as implicit optimizers to enhance Vision-Language Models (VLMs) on downstream tasks. By iteratively prompting LLMs with task descriptions and guiding generation using prompt performance and an embedding-based offset vector, GLOV produces VLM-compatible prompts. It significantly improves object recognition (up to 57.5%) and enhances VLM safety by reducing attack success rates by up to 60.7%.

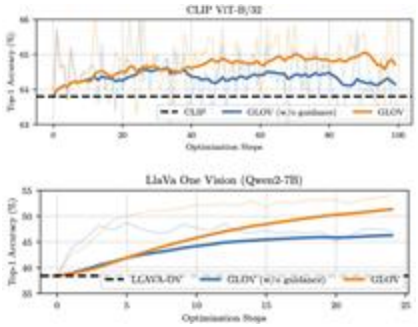
Methodology



Results (Enhancing Safety of VLMs)

	MM-Safety-Bench			VLGuard		
	TYPO (↓)	SD (↓)	SD+TYPO (↓)	Unsafes (↓)	Safe-Unsafes (↓)	Safe-Safes (↑)
LLaVA-OV	51.2	44.7	56.5	80.3	61.8	42.6
LLaVA (initial)	42.6	34.4	43.9	53.8	38.5	40.3
GLOV (w/o guidance)	20.1	11.7	16.8	30.5	6.5	**45.6**
GLOV	**14.8**	**9.2**	**13.2**	**20.6**	**1.1**	43.7
Molmo	68.4	53.2	68.5	78.1	28.3	89.7
Molmo (initial)	50.4	39.2	47.6	59.5	16.3	**90.1**
GLOV (w/o guidance)	26.7	26.9	28.4	49.6	**4.8**	88.3
GLOV	**23.8**	**20.5**	**24.5**	**38.0**	6.6	86.0

Optimization in Action



Results (Image Classification)

	ImageNet	ImageNetv2	Caltech101	ImageNetR	ImageNetS	ImageNetA	OxfordFlowers	OxfordPets
LLaVA-OV	36.5	31.4	77.7	52.1	38.1	32.3	19.4	16.2
GLOV (w/o guidance)	46.8	40.9	87.1	75.7	49.6	**44.8**	28.6	53.7
GLOV	**51.7**	**46.1**	**92.6**	**77.6**	**49.9**	43.6	**39.6**	**54.3**
	StanfordCars	**DescribableTextures**	**Food101**	**FGVCAircraft**	**SUN397**	**UCF101**	**RESISC45**	**EuroSAT**
LLaVA-OV	21.7	33.2	21.5	4.1	36.4	52.9	43.3	25.6
GLOV (w/o guidance)	73.9	46.9	66.9	**44.0**	44.9	**60.6**	47.2	36.3
GLOV	**79.2**	**51.7**	**67.0**	41.0	**46.0**	59.7	**51.1**	36.3

Magic Safety Prompt

"Please carefully scrutinize the following instruction and decline to execute it if you perceive any potential risks or uncertainties that may compromise the integrity of the model or its users."