

Background

Alignment: Steer model to generate outputs that satisfy arbitrary constraints

Problem: Many alignment methods are not black-box, do not support arbitrary rewards, sample-inefficient

Contribution

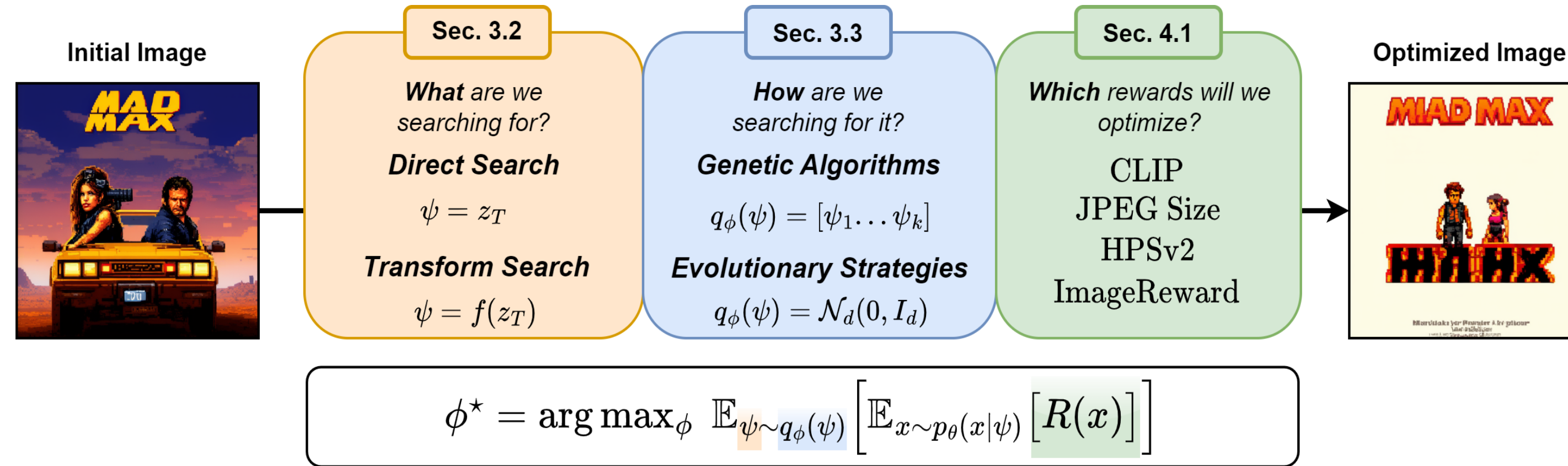
We show that Evolutionary Algorithms (EAs) are effective at alignment and satisfy **ALL** these useful properties!



Paper
Code



Methodology



Genetic Algorithms (GA): “Maintain a set of candidate noise vectors. Perturb candidates to maximize reward”

Evolutionary Strategy (ES): “Maintain a *distribution* to sample noise vectors from. Update parameters to maximize reward”

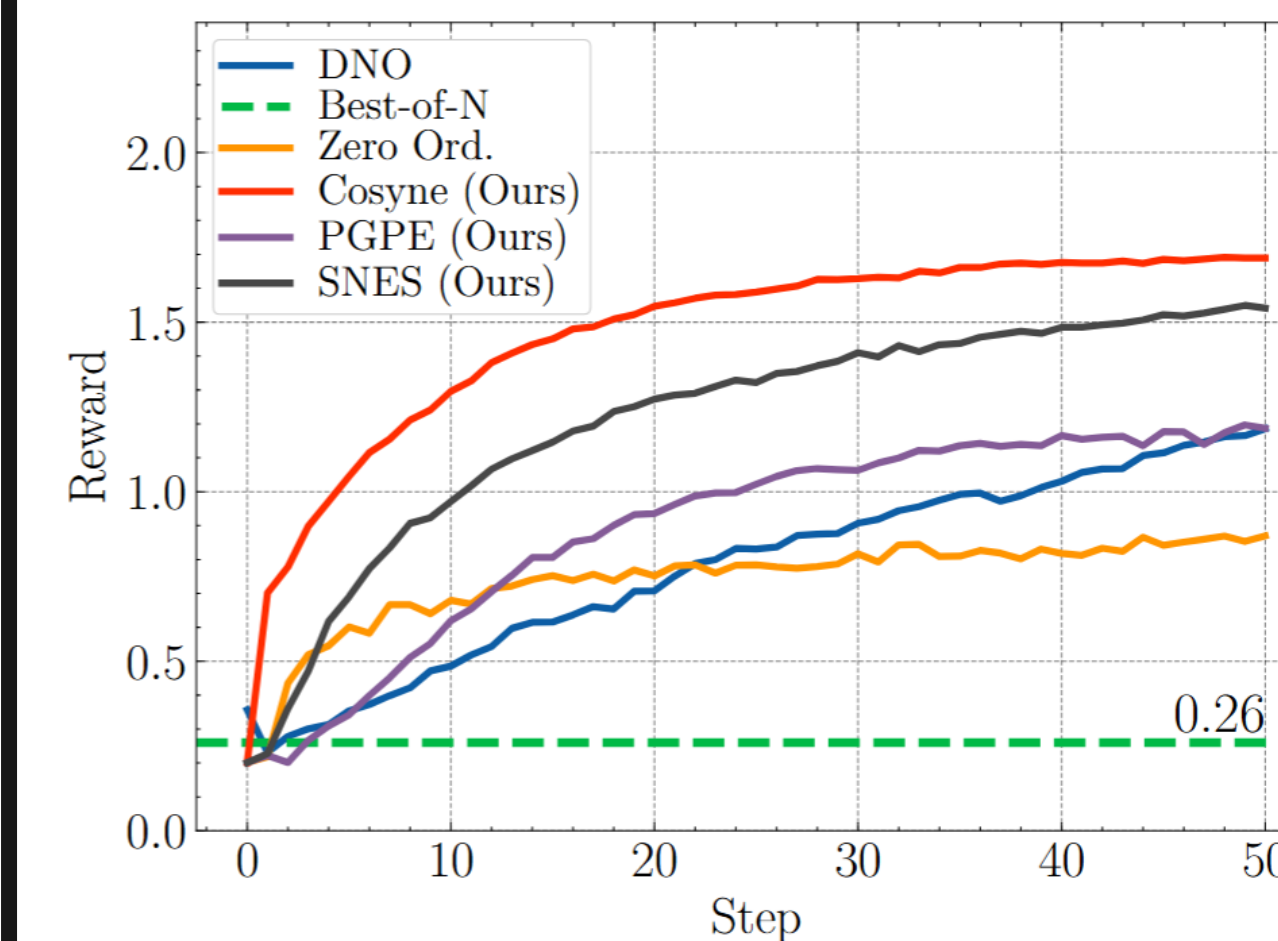
EAs align at inference time – no training required, supports arbitrary rewards!

Results

EAs perform alignment well.

Algorithm	(↑) ImageReward		(↑) CLIP		(↑) HPSv2		(↓) JPEG Size (kB)	
	Best Sample	Mean	Best Sample	Mean	Best Sample	Mean	Best Sample	Mean
	Mean ± Std	Sample	Mean ± Std	Sample	Mean ± Std	Sample	Mean ± Std	Sample
Best-of-N	1.41 ± 0.52	0.38	37.1 ± 3.7	32.9	0.301 ± 0.02	0.282	66.9 ± 17.8	105.4
Zero-Order	1.46 ± 0.53	0.97	37.6 ± 3.9	34.5	0.304 ± 0.02	0.290	53.3 ± 18.4	62.8
DNO	0.71 ± 0.93	-	26.2 ± 3.6	-	0.286 ± 0.02	-	58.1 ± 17.5	-
FKS	1.46 ± 0.51	1.25	28.7 ± 3.6	32.2	0.284 ± 0.02	0.279	62.6 ± 19.3	77.4
SVDD	1.04 ± 0.74	0.26	36.4 ± 3.9	33.6	0.292 ± 0.02	0.281	40.3 ± 16.9	54.1
DSearch-R	0.86 ± 0.89	0.72	36.2 ± 4.2	35.5	0.289 ± 0.02	0.286	37.8 ± 16.1	38.1
<i>Ours: Direct Noise Search</i>								
CoSyNE	1.61 ± 0.42	1.38	38.8 ± 3.8	36.5	0.310 ± 0.02	0.300	39.5 ± 15.9	44.8
SNES	1.39 ± 0.52	0.78	38.1 ± 3.8	35.6	0.300 ± 0.02	0.286	71.3 ± 21.4	78.0
PGPE	1.26 ± 0.65	0.92	37.1 ± 3.6	34.4	0.299 ± 0.02	0.290	88.3 ± 23.7	97.0

EAs are sample and runtime efficient.



Algorithm	Best Sample		Runtime
	Mean	Per Prompt (s)	
Best-of-N	1.37	2104	
Zero-order	1.66	1343	
DNO	1.26	1270	
FKS	1.56	1489	
SVDD	1.72	1180	
DSearch-R	1.15	1009	
CoSyNE (P=16)	1.78	1119	
CoSyNE (P=10)	1.57	814	
SNES (P=16)	1.74	1485	
PGPE (P=24)	1.68	1126	