Arena-T2I Hard

Benchmarking and Improving Faithfulness with Dependency-Aware Checklist

Yuanhao Ban, Tong Xie, Sohyun An, Yunqi Hong, Evan Frick
I-Hung Hsu, Wei-Lin Chiang, Ion Stoica, Cho-Jui Hsieh

Arena Intelligence Inc. and UCLA

NeurIPS 2026 Spotlight

Real users ask for many details at once

A real-user request from Arena-T2I-Hard

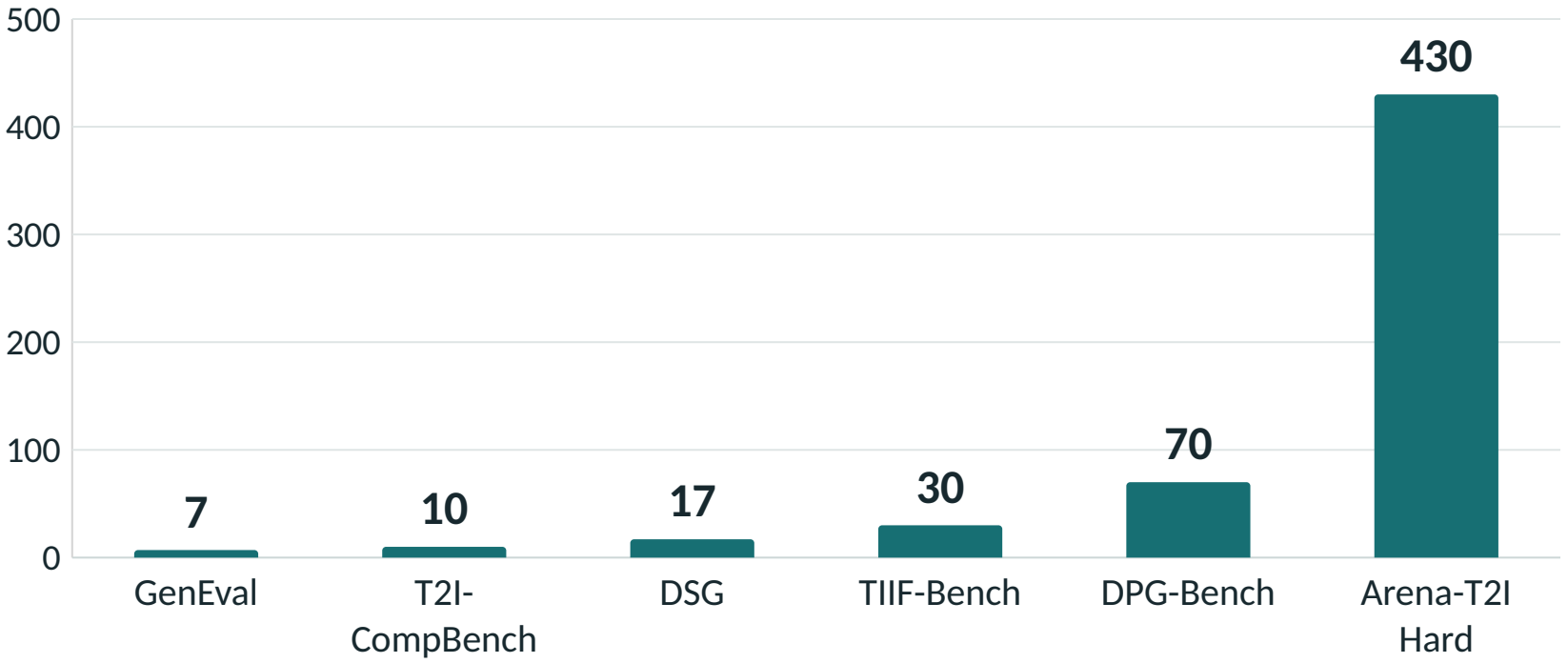
Create a hyper-stylized vector comic-book poster with five anthropomorphic monster trucks. Place “MONSTER” over 3D “TRUCKS”. Add speed lines, mud splatter, and checkered flags. Use a vibrant primary-color palette on a green background.

Object counts **Layout** **Style** **Exact text**

Even frontier image models miss requirements in complex real-user prompts.

Arena-T2I Hard uses complex real-user prompts

Average prompt length (words, approximate)



310

real-user prompts

~30

requirements per prompt

Faithfulness = satisfied checklist items / total items

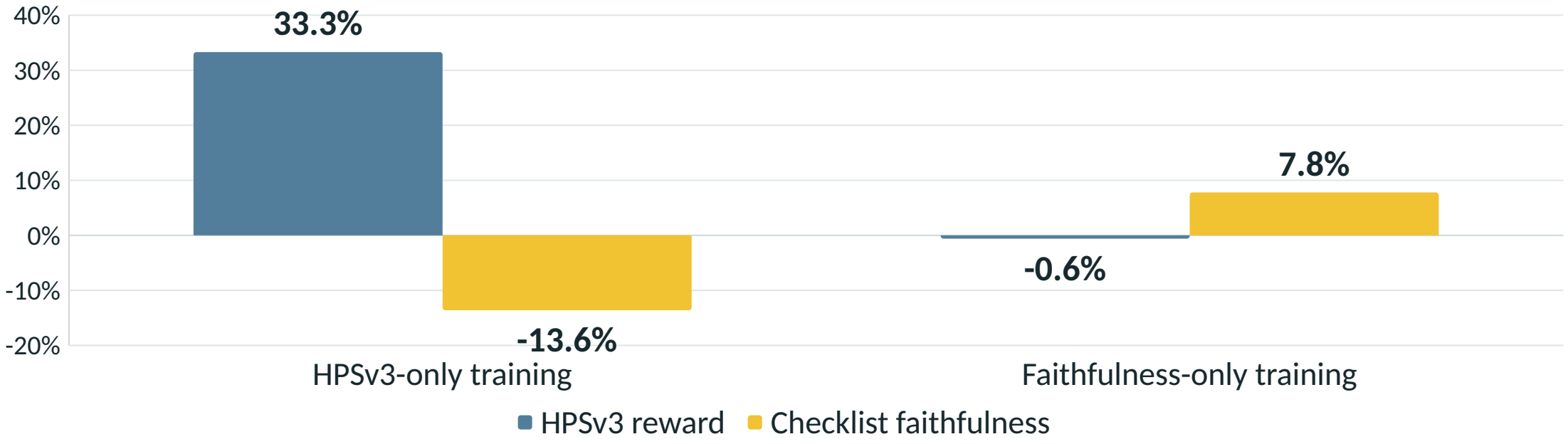
The strongest model reaches 85.5% faithfulness

Model	Arena-T2I Hard	DPG-Bench	DSG	Arena rank	Arena score
Gemini 3 Pro	**85.5%**	97.0%	94.6%	#3	1244 ± 4
Grok Imagine	**84.9%**	96.5%	93.4%	#8	1170 ± 4
GPT Image 1.5	**79.6%**	97.0%	94.2%	#4	1242 ± 4
Recraft V4	**78.7%**	95.6%	91.4%	#29	1109 ± 5
Wan 2.6	**76.8%**	95.4%	91.7%	#21	1132 ± 4
Gemini 2.5 Flash	**76.8%**	95.4%	92.1%	#14	1152 ± 3
GPT Image 1	**72.2%**	95.9%	93.8%	#27	1115 ± 3
Imagen 4 Ultra	**68.0%**	96.1%	92.8%	#17	1148 ± 4
Imagen 4	**65.9%**	94.7%	91.3%	#22	1130 ± 3
Hunyuan Image 3.0	**60.9%**	94.7%	83.7%	#15	1151 ± 3
Ideogram V3	**52.3%**	89.4%	85.5%	#42	1049 ± 4

Benchmark scores shown as percentages. Arena score includes reported uncertainty.

Faithfulness collapse

Preference-only training can weaken prompt following even as its reward increases.



SD3.5-Medium. Relative change from the initial checkpoint after 1,000 steps.

Dependency-aware checklists

Example request: “A blue chair”

Checklist item	Dependency	If the chair is missing
Is the chair present?	None	Not satisfied
Is the chair blue?	Chair must be present	No credit

Checklist satisfaction becomes a training reward

The model must satisfy individual requirements and respect their dependencies.

Pairwise evaluation on SD3.5-Medium

Each cell is the row method's win rate against the column method (%).

Row / column	Base	PickScore	HPSv3	Image Reward	Ensemble	Ours	DreamSync
Base	—	47.8	67.2	66.1	61.8	39.1	65.5
PickScore	52.2	—	72.9	71.7	65.4	40.2	65.8
HPSv3	32.9	27.2	—	45.3	40.6	22.7	45.7
ImageReward	33.9	28.3	54.7	—	43.5	22.5	50.8
Ensemble	38.2	34.6	59.4	56.5	—	28.3	52.7
Ours	**60.9**	**59.8**	**77.3**	**77.5**	**71.7**	**—**	**74.3**
DreamSync	34.5	34.2	54.3	49.2	47.3	25.7	—

Ours: checklist faithfulness + PickScore

Four-reward ensemble: PickScore + HPSv3 + ImageReward + UnifiedReward-2.0

Faithfulness needs an explicit training signal

01 Complex real-user prompts remain difficult

Even the strongest evaluated model reaches only 85.5% mean faithfulness.

02 Preference reward can rise while faithfulness falls

Dependency-aware checklist rewards explicitly target prompt requirements.

03 Checklist faithfulness + PickScore improves results

71.7% and 68.6% pairwise wins against the four-reward ensemble.

Paper: arxiv.org/abs/2606.31711