TIFA: Text-to-Image Faithfulness Evaluation with Question Answering

Yushi Hu[1], Benlin Liu[1], Jungo Kasai[1], Yizhong Wang[1], Mari Ostendorf[1], Ranjay Krishna[1,2], Noah A. Smith[1,2]

University of Washington[1] Allen Institute for AI[2]

Background: we need a measure for text-to-image faithfulness

- Text-to-Image model does not follow user inputs
- Prior papers just show good demos
- CLIP is problematic and does not align with human judgment

Examples:

Text Input: A person sitting on a horse in air over gate in grass with people and trees in background.

CLIP: 24.3

TIFA 71.4

CLIP: 21.3

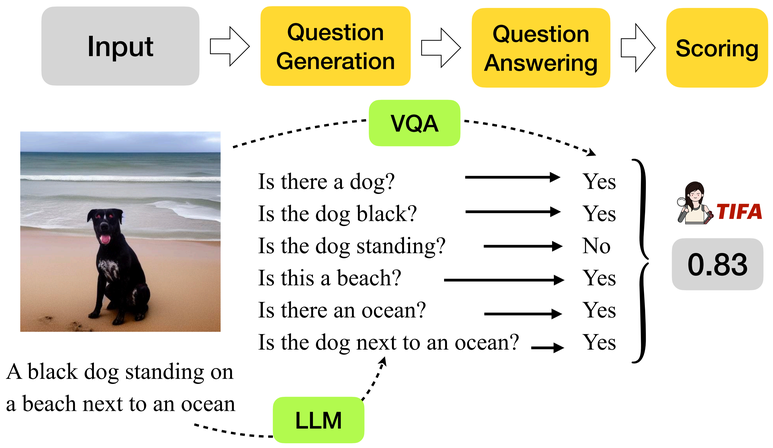
TIFA 100.0

Our solution

Use powerful **LLMs** and **VLMs** to analyze images

1. Parse Texts and Generate questions with LLMs
2. VQA with VLMs (BLIP, PaLI, GPT-4, …)
3. Compute faithfulness by VQA accuracy

TIFA's advantage:
Accurate, Interpretable, Fine-Grained, Modular



Human Annotation

We collect human Likert scale (1-5) annotation for image-text alignment

TIFA has a much higher correlation with human judgments!

	Spearman's ρ	Kendall's τ
Caption-Based		
BLEU-4	18.3	18.8
ROUGE-L	32.9	24.5
METEOR	34.0	27.4
SPICE	32.8	23.2
CLIPScore	33.2	23.1
Ours		
TIFA (VILT)	49.3	38.2
TIFA (OFA)	49.6	37.2
TIFA (GIT)	54.5	42.6
TIFA (BLIP-2)	55.9	43.6
TIFA (mPLUG)	**59.7**	**47.2**

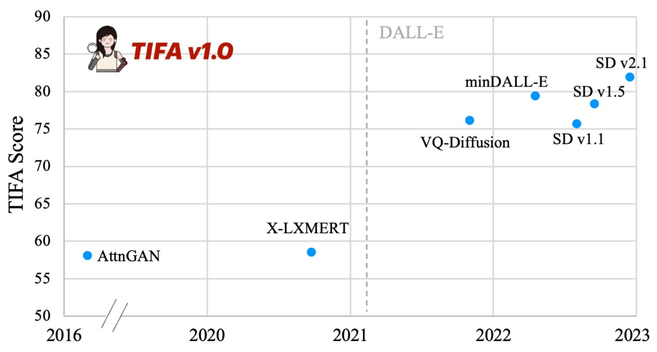
TIFA v1.0 Benchmark

We collect a benchmark for T2I Faithfulness

- 4,000 diverse prompts
- 25,000 questions
- 12 Skill Categories (objects, attribute, relations, counting, …)
- 4,550 distinct visual elements
- SOTA VQA models like BLIP-2

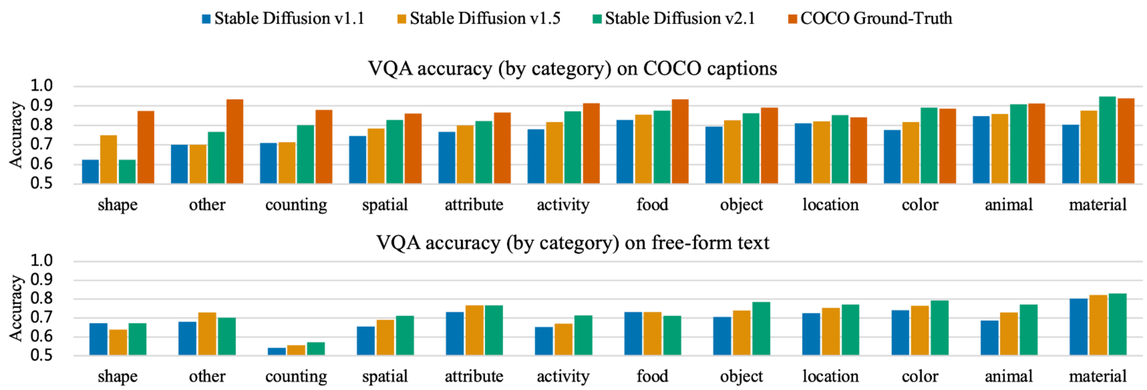
"a panda bear with aviator glass"
"a fox in the style of starry night"
"one cat and two dogs sitting on grass"
…

Evaluation of Recent T2I models



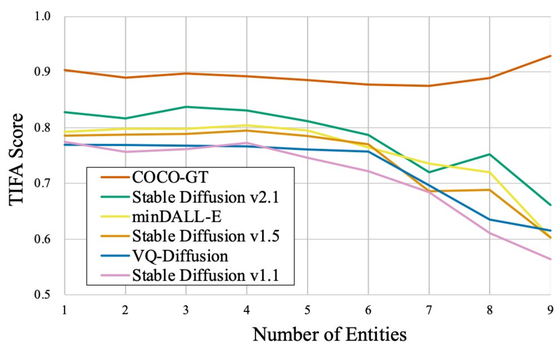
- Models are improving!
- SDXL got 85.27

What are T2I models struggling on?



- **Bad at shape, counting, spatial relations, abstract art notions**
- Good at material, animal, color
- Generate from free-form texts is harder than from image captions

Compositionality is hard!



- **TIFA** score drops rapidly when more entities are added to the text prompts

Code Data

Contact

yushihu@uw.edu

@huyushi98