TIFA: Text-to-Image Faithfulness Evaluation with Question Answering
Yushi Hu¹, Benlin Liu¹, Jungo Kasai¹, Yizhong Wang¹, Mari Ostendorf¹, Ranjay Krishna¹,², Noah A. Smith¹,²
University of Washington¹ Allen Institute for AI²

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.

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!

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

Evaluation of Recent T2I models

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

Examples:
- "a panda bear with aviator glasses"
- "a fox in the style of starry night"
- "one cat and two dogs sitting on grass"

Code Data Contact
yushihu@uw.edu @huyushi98