Imagen Editor & EditBench

Advancing and Evaluating Text-Guide Image Inpainting

Su Wang*, Chitwan Saharia*, Ceslee Montgomery*,
Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut,
Jason Baldridge,
Mohammad Norouzi†, Peter Anderson†, William Chan†
Google Research, Brain Team

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Text-Guided Image Inpainting

Task: Given an **image**, a **mask**ed area, and a **text** instruction → Edit the masked area according to text while keeping the unmasked area intact.

Use cases:

- **Image editing**: Tinkering a raw generated/retrieved image to fit the user's vision better.
- **Data synthesis**: Creating training/evaluation data for text-image modeling at scale with balanced content distribution (addressing the "long tail").
  `(Example: in ROSIE, we generated diverse environments to teach robots to learn navigation and manipulation better!)`

Text-Guided Image Inpainting

Evaluation: When we say Model A performs better/worse than Model B, what do we mean?

Existing metrics are

- Reasonably reliable at model/dataset-level, while leaving **rooms of improvement at instance-level** (e.g. CLIP-based metrics, BLEU/METEOR/SPICE).
- Need to be made **more informative to answer the question “then how do we improve?”** (e.g. human judgments on “does this image match this text?”).

Google

In this work …

We present a text-guided image inpainting model — **Imagen Editor** — that is steps ahead of the state-of-the-art competition.

… and to back up our claim, we propose a systematic and fine-grained evaluation benchmark — **EditBench** — to stack Imagen Editor against competition in comprehensive experimentation.

The benchmark is also an effort to push for more informative, semantically granular text-image alignment evaluation *beyond the common "wall of pretty images + hard-to-interpret auto/human evals" formula.*

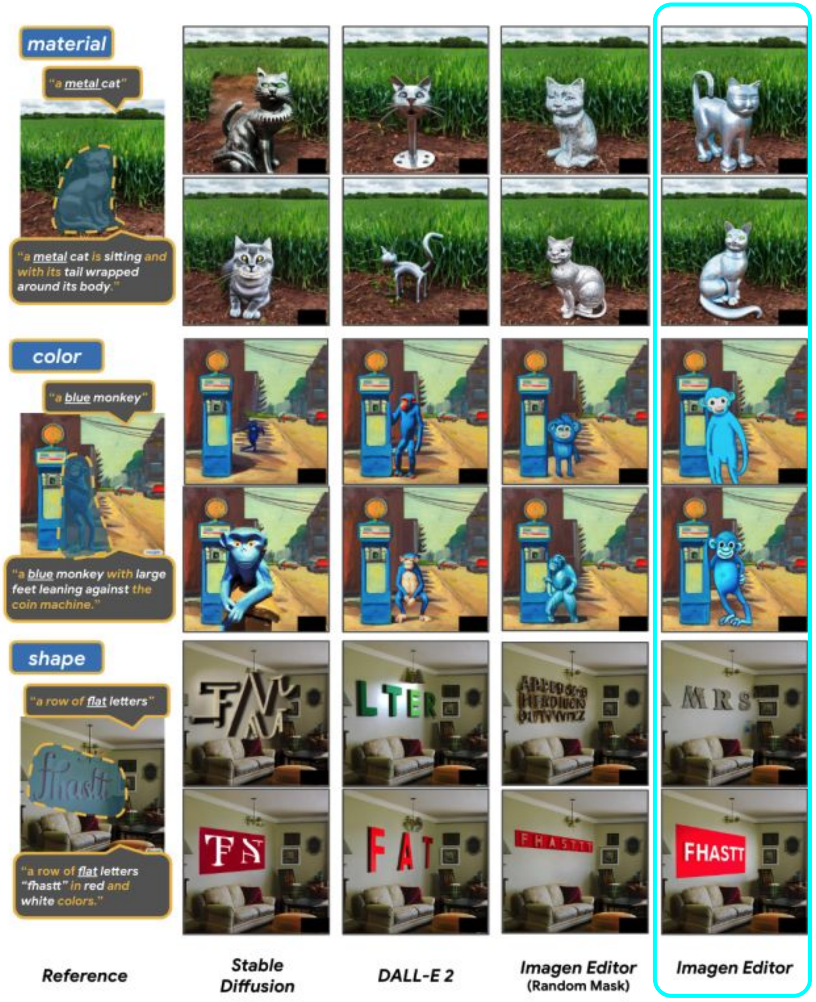


Imagen Editor

A cascaded diffusion model built on our text→image generator **Imagen**.

High-resolution Conditioning.
All cascading stages are conditioned on the full resolution image (the editing target).

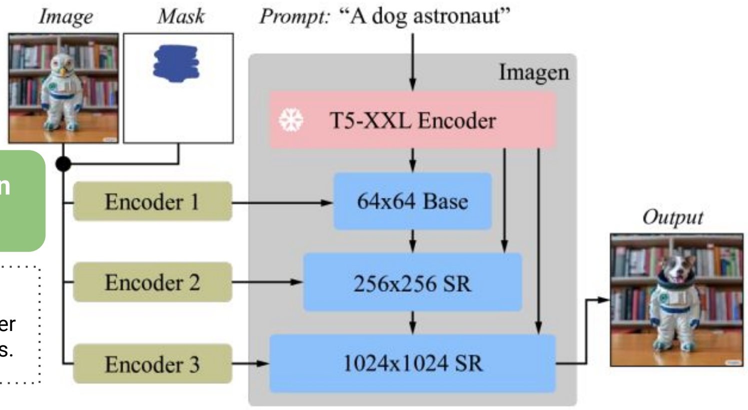
Object-oriented Masking. Denoising mask created by powerful object detection model on-the-fly.

Classifier-Free Guidance. Biasing output towards more faithful text-image alignment.

NOTE: This is **in addition** to the standard training techniques (rather than replacing) such as random masking with boxes / strokes, uncropping, flipping, etc.

High-resolution Conditioning

Better inpainting result across a wider range of resolutions.



Object-oriented Masking

Greatly mitigates paint-over and artifacts.

Classifier-Free Guidance

Follows the text instruction better.

EditBench

A collection of 240 rich annotated evaluation items (50:50 generated vs. natural image ratio) that features

- **Fine-grained semantic categories** along three dimensions
  - `Attributes`: `{material, color, shape, size, count}`
  - `Objects`: `{common, rare, text rendering}`
  - `Scenes`: `{indoor, outdoor, realistic, painting}`

- **Text prompt types**
  - `Full-image`: `describes the entire image.`
  - `Mask-simple`: `describes only the main object/attribute in the masked area.`
  - `Mask-rich`: `Also only targets the masked area but more richly descriptive.`

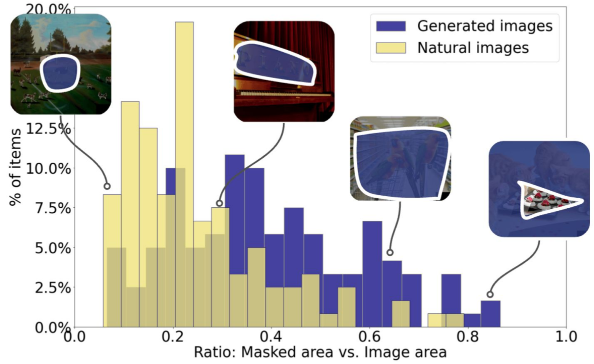
- Covers a wide **range of mask sizes**

*

Mask Simple
"A metal cat"

Mask Rich
"A metal cat **is sitting and with its tail wrapped around its body**"

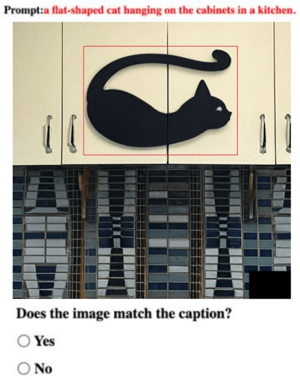
Full Image
"A metal cat **sitting in the middle of a farm field.**



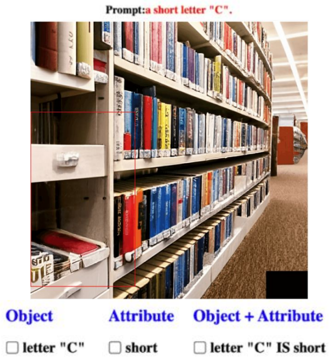
Google

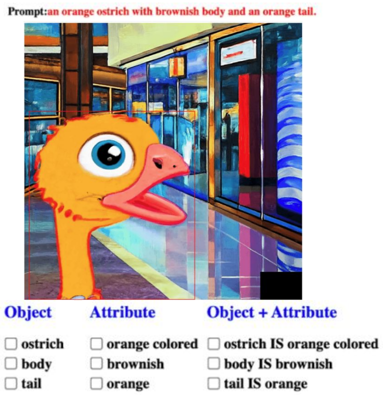
EditBench

Leveraging the rich annotation, we elicit human judgment from different angles in a fine-grained way.

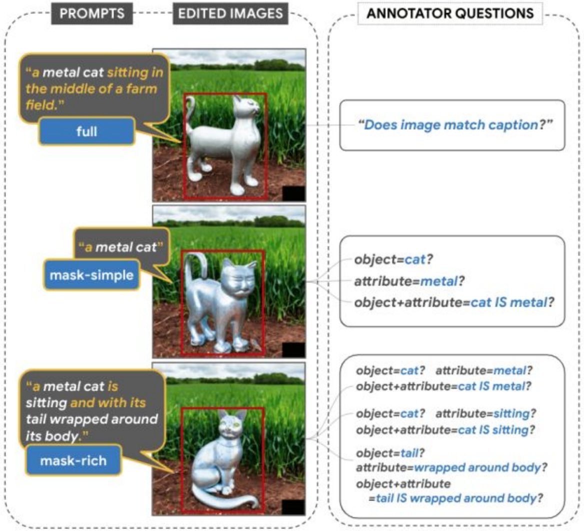


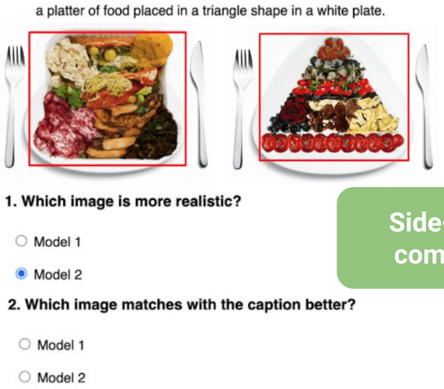
General matching question





Detailed object/attribute questions





Side-by-side comparison

Google

Experimentation

We compare

- **Stable Diffusion** (v1.5, latest at the time of pub)
- **DALL-E 2**
- **Imagen Editor baseline** (no object masking)
- **Imagen Editor**

with

- **Automatic Evaluation** (standard method sa. CLIPScore)
- **Human Evaluation** (with EditBench)

Result – Automatic Evaluation

Experiment 1. Scoring directly

- **CLIPScore**: Text-Image embedding similarity.
- **CLIP R-Prec**: Retrieval precision of edited image for ground truth text from a set of 100.
- **NIMA**: Image quality assessment based on human perceptual quality and aesthetics.

	SD	DL2	IM_{RM}	IM	Ref.
CLIPScore (↑)					
T2I	29.7	29.1	29.6	**31.5**	31.0
I2I	74.9	76.1	75.8	**76.6**	-
T2I+I2I	52.3	52.6	53.1	**53.6**	-
CLIP-R-Prec (↑)	96.5	95.3	95.0	**98.6**	$\boxed{99.3}$
NIMA (↑)	4.44	4.33	4.56	**4.63**	$\boxed{4.89}$

Experiment 2. Model selection

Percentage agreement between CLIPScore metrics and human judgments when picking the best image out of two model-produced images for the same prompt.

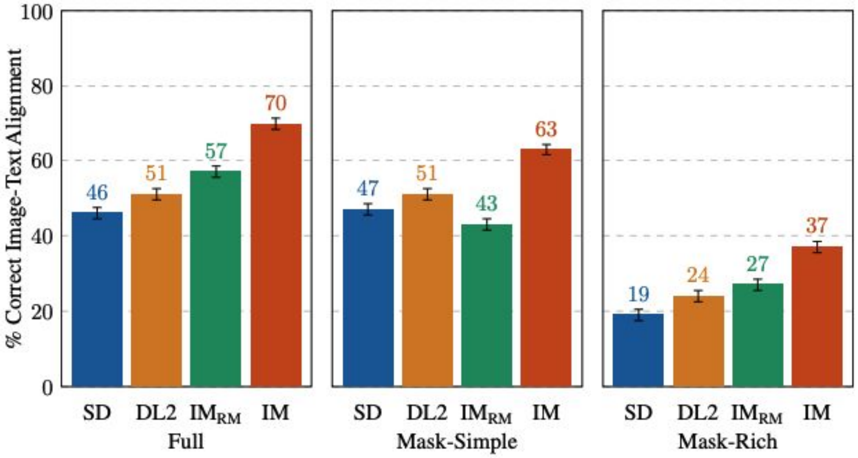
Prompt	Image	T2I	I2I	T2I+I2I	R-Prec	Rand
Full	Full	**70.1**	58.6	66.8	53.3	50.0
Full	Crop	68.1	55.8	62.4	57.7	50.0
Mask-Simple	Full	73.8	53.1	63.2	72.0	50.0
Mask-Simple	Crop	**76.0**	55.3	66.4	71.0	50.0
Mask-Rich	Full	66.7	55.2	63.4	62.3	50.0
Mask-Rich	Crop	**68.4**	56.4	64.1	63.3	50.0

Result – Human Evaluation with EditBench

Experiment 1. Single-image evaluation

probes a model with three types of prompts:

- **Full prompts** that describe the entire image – we elicit binary answers to the question Does the image match the caption?.
- **Mask-Simple prompts** describe the masked area only and involve a single attribute-object pair – we check if the object and attribute are properly rendered, as well as whether they are bound to each other correctly (e.g. for red cat, a white cat on a red table would be an incorrect binding).
- **Mask-Rich prompts** extend mask-simple to 3 or more object-attribute pairs.



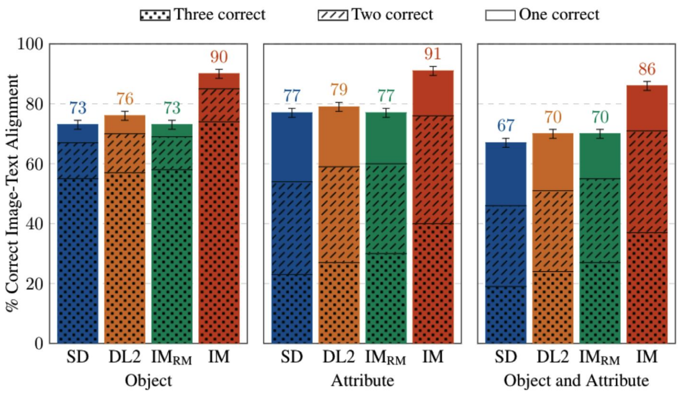
For mask-simple/rich prompts, text-image alignment is only counted as correct if the edited image correctly includes every attribute and object specified in the prompt, including the correct attribute binding (setting a very high bar for correctness)

Google

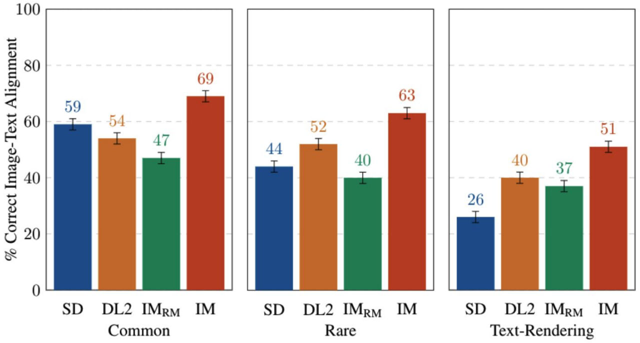
Result – Human Evaluation with EditBench

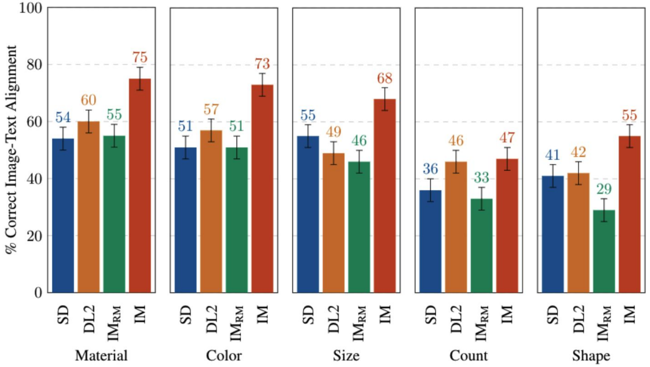
Experiment 1. Single-image evaluation (con'td)

- Breakdown by object types (Top)
- Breakdown by attribute types (Bottom)



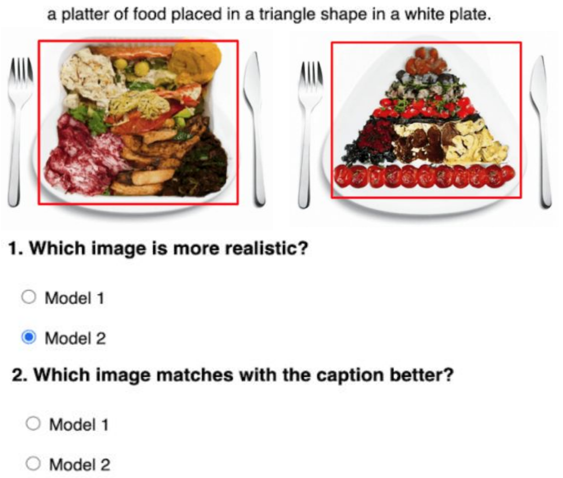
A breakdown of the percentage of correctly aligned object, attribute, and object-attribute binding (with the prompt type Mask-Rich)

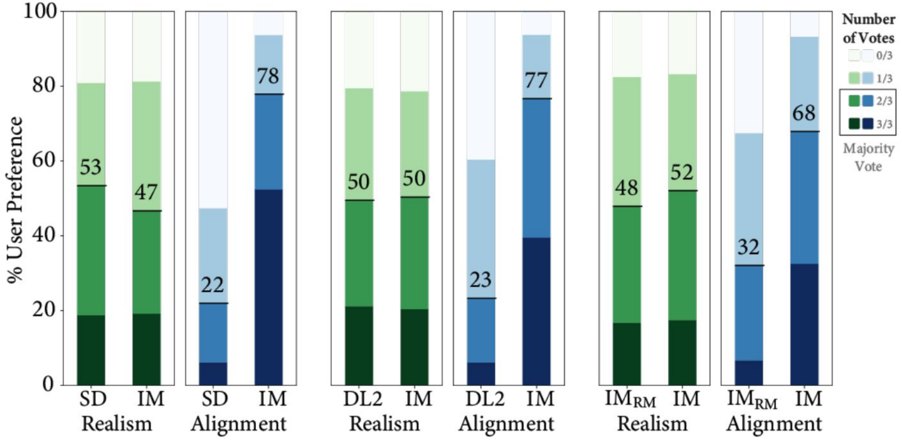




Result – Human Evaluation with EditBench

Experiment 2. Side-by-side evaluation



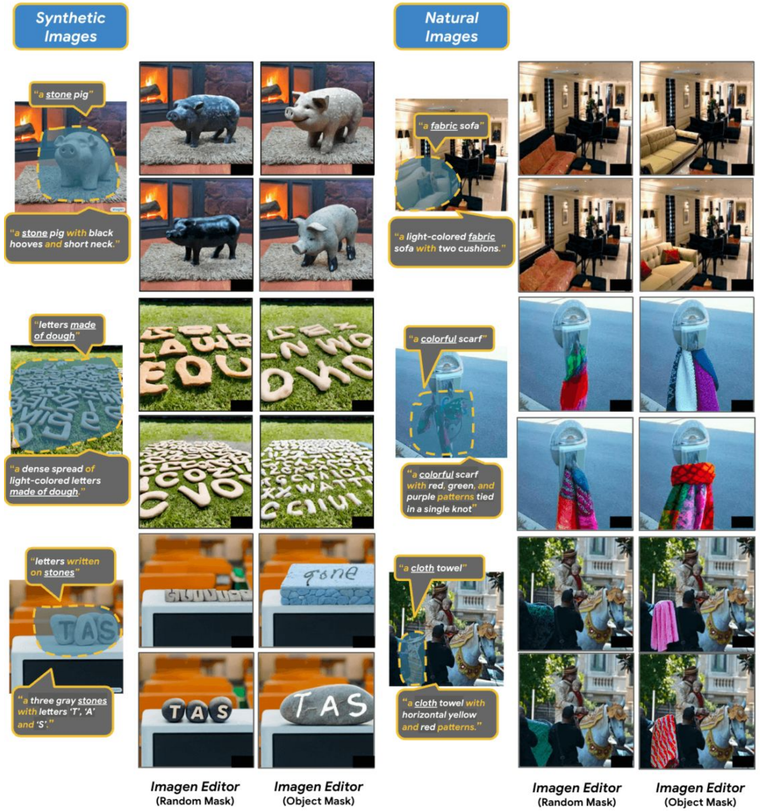


Result – Human Evaluation with EditBench

Experiment 3. Qualitative analysis

Object-oriented masking greatly improve models' ability to faithfully follow text instructions. Specifically, include objects/attributes mentioned correctly.

This is also quantified in our breakdown of the percentage of objects/attributes/bindings correctly rendered.

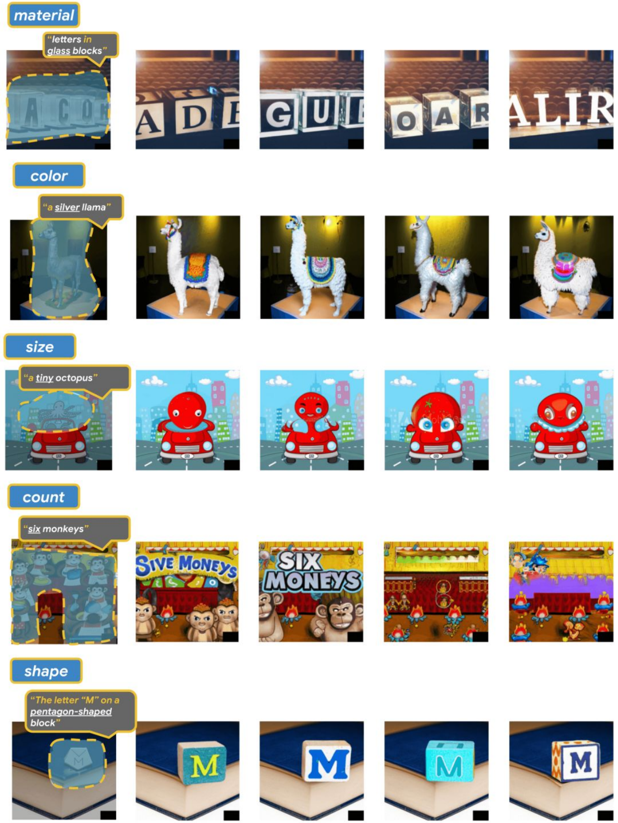


Result – Human Evaluation with EditBench

Experiment 3. Qualitative analysis (cont'd)

Some semantic categories are harder than others.

- First-order semantic properties such as material, color, size, etc. are easier.
- Less abstract properties such as color are easier than more abstract ones such as shape.
- Higher-order properties such as count are harder.



Summary of Findings

- **Imagen Editor** is a state-of-the-art text-guided image inpainting model, based on our automatic and human evaluation results.
- The human evaluation results produced with our **EditBench** benchmark are fine-grained, interpretable, and informative for the further development of text-guided image editors.
- Our findings illustrate a promising and functionally useful direction for the evaluation of text-image models. We hope it seeds and leads to more innovations in the space.

Thank you!