A Appendix

A.1 Dataset Supplementary Materials

1. Dataset documentation, metadata, and download instructions: `anonymous`.
2. Intended uses: we hope SeeTRUE will be used by researchers to evaluate image-text matching models.
3. Author statement: We bear all responsibility in case of violation of right in using our benchmark.
4. Each dataset's license is described below. Our additional human annotations and geenrated images are licensed under CC-BY 4.0 license `https://creativecommons.org/licenses/by/4.0/legalcode`.
5. Hosting & preservation: the dataset will be hosted in Huggingface Datasets, accessible and available for open research.
6. SeeTRUE fields are presented in table 5.

Additional licensing details: we do publish the datasets we annotated in this work, and do not re-publish the existing SNLI-VE and Winoground datasets. Full licenses:

1. MS COCO [28]: `https://cocodataset.org/#termsofuse`
2. EditBench [25]: `https://research.google/resources/datasets/editbench/`, `https://www.apache.org/licenses/LICENSE-2.0`
3. DrawBench [5]: `https://imagen.research.google/`, `https://docs.google.com/spreadsheets/d/1y7nAbmR4FREi6npB1u-Bo3GFdwdOPYJc617rBOxIRHY/edit#gid=0`
4. Pick-a-Pick [29]: `https://huggingface.co/datasets/yuvalkirstain/pickapic_v1`
5. SNLI-VE [27]: `https://github.com/necla-ml/SNLI-VE`
6. Winoground [24]: `https://huggingface.co/datasets/facebook/winoground`

Table 5: SeeTRUE Rows Examples

image	text	label	original_dataset_id	dataset_source
img1	A zebra to the right of a fire hydrant.	0	text_133_image_1228	drawbench
img2	A group of people standing next to bags of luggage.	1	text_105_image_1377	coco_t2i
img3	a tiny figurine is surrounded by cell phones on a table.	1	3786	editbench

A.2 Human Annotation Process

Figure 7: Annotation interface for determining whether a given image-text pair are aligned.

In order to provide reliable human labels for our datasets, we conducted an annotation process using the SeeTRUE platform. The process comprised several steps, including setting qualification requirements, providing instructions, and evaluating annotator agreement.

We set the basic requirements for our annotation task as follows: a percentage of approved assignments above 98%, more than 5,000 approved HITs, and annotator locations limited to the US, UK, Australia, or New Zealand. We selected 5 examples from our dataset for a qualification test and screened the annotators' results. fig. 7 displays a sample of the Mechanical Turk user interface. The payment for the crowd-workers was 15-18 USD hourly.

The instructions provided were as follows:

Evaluate the given image and text to determine if they match, selecting either "Yes" or "No". Some images may be synthetically generated by a text-to-image model. To assess the match, mentally generate a textual description for the image (no need to write it down) and compare this generated description to the given text. If the descriptions closely resemble each other, mark "Yes". If not, mark "No" and provide feedback on the specific issue causing the misalignment, focusing on the primary issue if multiple misalignments are present. If you encounter an image or text that may be offensive due to bias, race, NSFW content, etc., mark the checkbox to indicate this issue.

Full agreement metrics are presented in table 6. As shown in the table, the percentage of cases where all annotators agreed and the Fleiss-Kappa scores vary across the datasets, with COCO-Con exhibiting the highest level of agreement and Drawbench the lowest. These differences highlight the varying levels of complexity within the datasets.

Table 6: Agreement metrics for different datasets.

Dataset	Full	Drawbench	COCO t2i	COCO-Con	PickaPic-Con
# Items	8,527	1,968	2,586	1,992	1,981
% all agreed	80	76	78	86	77
Fless-Kappa	0.72	0.66	0.68	0.81	0.69

A.3 Comparing VQ^2 variants

We have experimented with different variants of the VQ^2 pipeline.

Table 7: Comparing VQ^2 variants on all EditBench categories

Method	Extended answer candidates	Models used for validation	**EditBench** Full	Color	Count	Material	Shape	Size
VQ^2 (NLI)	✗	VQA & NLI				66.8		
VQ^2 (NLI)	✓	VQA & NLI						
VQ^2 (PaLI comparison)	✓	VQA & VQA						
VQ^2 (yes/no)	✗	VQA						
VQ^2 (yes/no)	✓	VQA						

A.4 Reproducibility

To fine-tune BLIP2, we adjust only the Q-former parameters of the model using the Adam optimizer. We train the model for two epochs and designate 10% of the training set as a validation set for early stopping and use learning rate selection between {1e-5, 5e-5}. A single training took 5 hours on a linux server with one A6000 GPU. All experiments took <2 days.

Zero-shot VQ^2: For 10,000 text-image pairs, the inference time of every step is as follows. Answer candidate generation: when using extended answer candidates – about 1 day. Otherwise, 12 hours. Question generation and filtering: When using extended answer candidates, about 2 days, otherwise, 1 day. The last step only takes a few minutes.