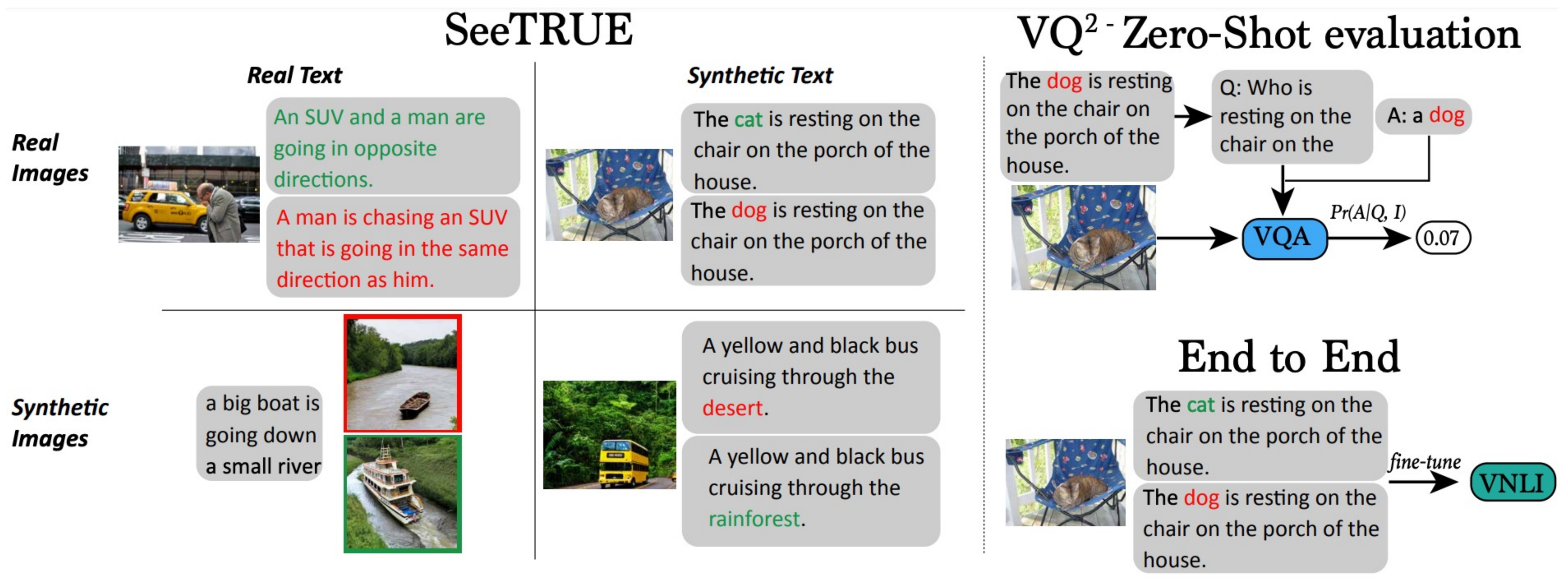
What You See is What You Read? Improving Text-Image Alignment Evaluation

Michal Yarom*, Yonatan Bitton*, Soravit "Beer" Changpinyo, Roee Aharoni,
Jonathan Herzig, Oran Lang, Eran Ofek, Idan Szpektor

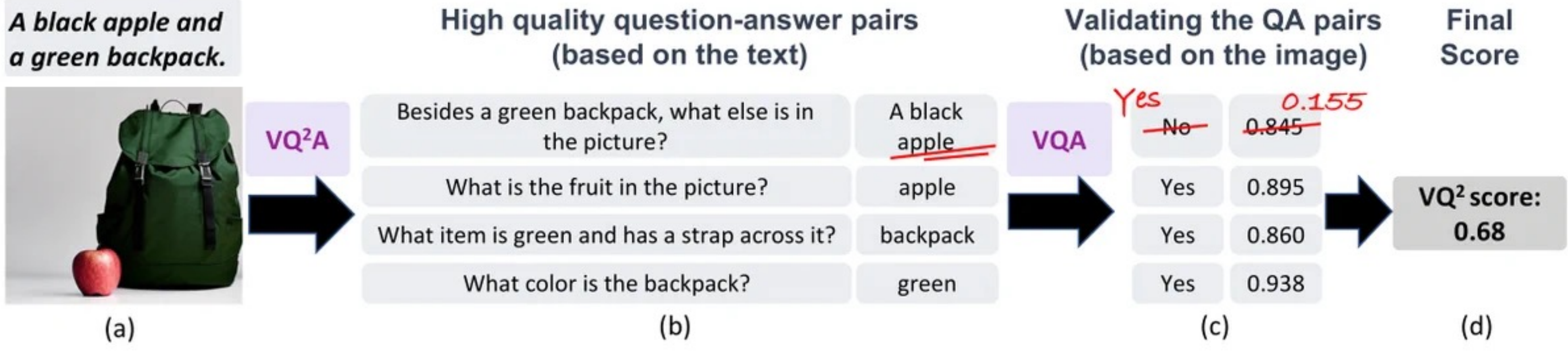
1. What You See is What You Read?

- Focusing on image-text alignment, we introduce SeeTRUE, a comprehensive benchmark, and two effective methods: a zero-shot VQA-based approach and a synthetically-trained, fine-tuned model, both enhancing alignment tasks and text-to-image reranking



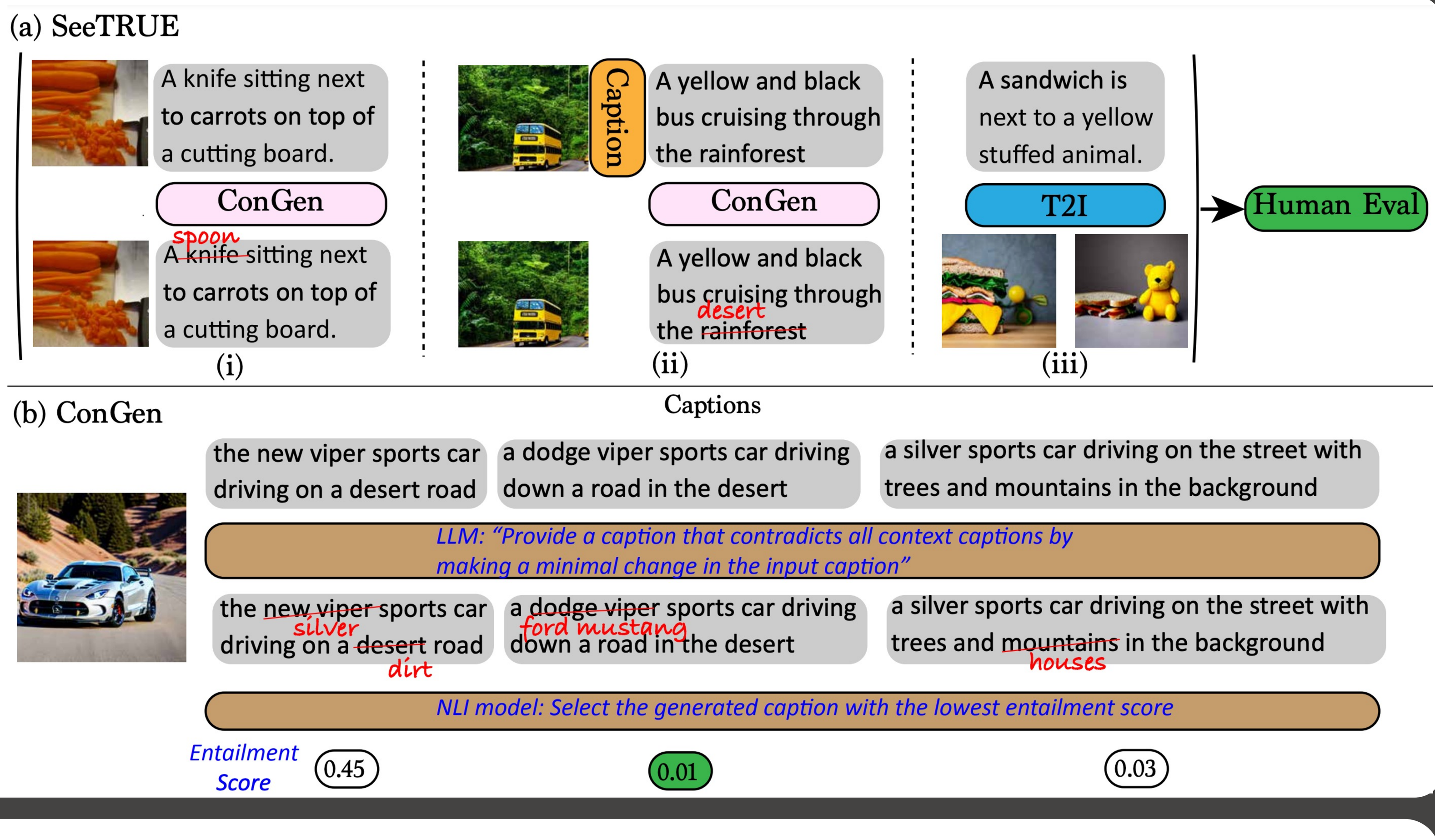
2. The VQ^2 Method

- Utilizes question generation and visual question answering
- Creates questions related to the text, ensuring the correct answer is obtained when asking these questions with the provided image



3. Visual Entailment Data Generation

- Generating data using text-to-image (t2i) models, image-to-text (i2t) models, large language models (LLMs), and natural language inference (NLI)
- Including a mix of natural and synthetic images, captions, and prompts
- ConGen: Generating Contradicting Captions by Prompting LLMs



4. Main Experiments

- Our methods surpass prior approaches in various text-image alignment tasks
- Significant improvements in challenging cases involving complex composition or synthetically generated images
- State-of-the-art results on the challenging Winoground dataset

	Text & Images	Real + Real		Real + Synthetic			Synthetic + Real	Synthetic + Synthetic	Avg.
	Model	SNLI-VE	Winoground	DrawBench	EditBench	COCO t2i	COCO-Con	PickaPic-Con	
zero-shot	CLIP RN50x64	66.6	53.6	59.2	67.1	58.8	71.1	66.8	63.3
	CLIP ViT-L14	65.8	53.3	60.5	62.1	58.8	70.7	66.8	62.6
	COCA ViT-L14	68.5	53.1	67.4	66.3	62.1	74.2	68.1	65.7
	COCA ViT-L14 (f.t on COCO)	70	53.1	66.2	68.3	66.2	76.5	67.2	66.8
	BLIP	75.2	58.2	60.5	68	70.7	84.2	76.6	70.5
	BLIP2	76.4	56.9	58.5	67.5	66.9	84.3	76.9	69.6
	BLIP 2 (f.t. COCO)	75.9	60	65.7	70	73.3	85.8	78	72.7
	PaLI	65.4	53.6	60.2	56.7	53.3	65.5	60.5	59.3
	TIFA	–	58.0	73.4	67.8	72.0	–	–	–
	VQ^2 (Ours)	88.0	**63.5**	82.6	73.6	**83.4**	87.1	81.7	**80.0**
f.t. snli-ve	OFA Large	80.5	53.3	77.6	70.9	67.5	75.4	69.5	70.7
	BLIP2	82.3	58.5	64.3	58.7	60.5	82.6	66.9	67.7
	PaLI	**95.1**	61.7	82.8	65.5	77.7	**91.2**	83.7	79.7
	PaLI + Synthetic Data	94.2	61.8	**86.8**	**77.2**	83.2	91	**85.9**	**82.9**
	Avg(VQ^2, PaLI+Syn)	93.9	**63.5**	**87.8**	**78.4**	**85.1**	**93**	**87.3**	**84.1**

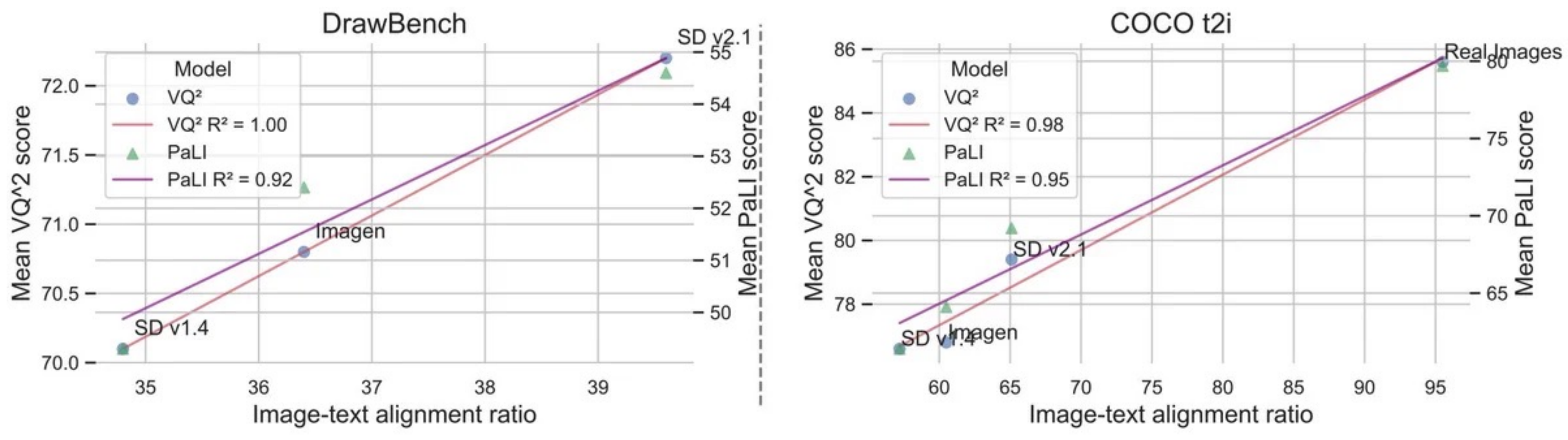
5. Contradiction Generation

- Our VQ^2 method detects inconsistencies between images and text by pinpointing question-answer pairs with the lowest VQA scores, proving effective across multiple datasets.

(a) "the orange lollipop is sad and the red lollipop is surprised" Q: What is the orange lollipop feeling? A: sad
(1) Winoground

(b) "Someone in a blue hat standing on a snowy hill" Q: What is the person wearing? A: blue hat
(2) CocoCon

(c) "A black apple and a green backpack" Q: What color is the apple? A: black
(3) DrawBench



6. Comparing Generative Models

- VQ^2 and VNLI scores are highly correlated with human ranking in evaluating text-to-image models
- Offers a way to evaluate dataset difficulty
- Revealing DrawBench as a harder dataset compared to COCO-t2i

7. Reranking Using Alignment Assessment

- Reranking image candidates - DrawBench and COCO-t2i
- VQ^2 and VNLI consistently achieves higher quality scores compared to CLIP
- Showcasing the potential in enhancing text-to-image systems

Dataset	Model	Random	CLIP	PaLI	VQ^2
COCO t2i	SD 1.4	68.6	74.6	88.2	86.4
	SD 2.1	71.3	81.2	84.5	87.3
DrawBench	SD 1.4	66.7	77.4	77.4	87.1
	SD 2.1	59.0	78.0	87.0	82.0

A brown and white cat is in a suitcase

VQ^2 score: 0.878 (1st)	VQ^2 score: 0.846 (2nd)	VQ^2 score: 0.731 (3rd)	VQ^2 score: 0.717 (4th)
PaLI ft. score: 0.992 (1st)	PaLI ft. score: 0.992 (2nd)	PaLI ft. score: 0.803 (3rd)	PaLI ft. score: 0.437 (4th)
CLIP similarity: 0.236 (4th)	CLIP similarity: 0.238 (3rd)	CLIP similarity: 0.253 (1st)	CLIP similarity: 0.25 (2nd)