eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grounding Language Models for Compositional and Spatial Reasoning

Master Thesis: Master in Language Analysis and Processing

Author: Julen Etxaniz

Advisors: Oier Lopez de Lacalle & Aitor Soroa

Departments: Computer Systems and Languages & Computational Science and Artificial Intelligence

Date: October 2022

1 Introduction

- Language models (LMs) have impressive capabilities in many tasks
- Pre-train and fine-tune to avoid training from scratch
- However, LMs criticized for lack of meaning
- Grounding necessary for human-like language understanding
- Compositional and spatial reasoning are challenging for LMs
- Vision and Language Models (VLMs) are pre-trained on image-text pairs
- VLMs better than LMs, but still struggle to ground spatial concepts
- Text-to-image diffusion models have some visual reasoning skills
- Diffusion models struggle to ground composition
- Mistakes: attribute leakage, interchanged attributes, missing objects

1.1 Objectives

- Three main objectives and sections
- Improve the state-of-the-art in compositional reasoning
    - Winoground zero-shot models not much better than chance, far from humans
    - We extend experiments with more pre-trained and fine-tuned models
- Perform zero-shot experiments in spatial reasoning
    - VSR fine-tuning large performance gap between models and humans
    - We extend experiments with zero-shot fine-tuned on NLVR2
- Investigate the use of synthetic datasets to overcome the lack of data
    - Text-to-image Stable Diffusion to generate images from Winoground captions
    - Image captioning models to generate captions for Winoground images
    - Image retrieval systems to retrieve images of interest from a large dataset

2 Winoground Zero-shot Experiments

- The Winoground paper only included zero-shot experiments with pre-trained models
- We test more pre-trained models and get better results
- We test models fine-tuned for specific tasks such as image-text retrieval and visual reasoning
- We prove that fine-tuning helps to obtain better results than previous experiments

2.1 Winogrond Dataset

- 400 test examples to probe visio-linguistic compositional reasoning
- Each one contains two images and captions, the goal is to match them
- Both captions contain the same words in a different order
- All examples are labelled with linguistic tags. 65 total, 3 main groups:
  - Object swaps consist in swapping noun phrases that refer to objects.
  - Relation swaps reorder words that refer to objects such as verbs, adjectives…
  - Both swaps involve changing both relations and objects.
- Linguistic swap independent: 1 or 2 main predicates
- Some examples have visual tags:
  - Pragmatics tag includes images that need to be interpreted non-literally
  - Series tag contains examples where both images come from the same photo series
  - Symbolic tag represents that the images include a symbolic representation

(a) [some plants] surrounding [a lightbulb]

(b) [a lightbulb] surrounding [some plants]

Object

(c) a [brown] dog is on a [white] couch

(d) a [white] dog is on a [brown] couch

Relation

(e) [circular] food on [heart-shaped] wood

(f) [heart-shaped] food on [circular] wood

Relation

Figure 3.1: Examples from the Winoground dataset for the swap-dependent linguistic tags *Object*, *Relation* and *Relation* from left to right. They are additionally tagged with 1 main predicate.

(a) there is [a mug] in [some grass]

(b) there is [some grass] in [a mug]

Object

(c) a person [sits] and a dog [stands]

(d) a person [stands] and a dog [sits]

Relation

(e) it's a [fire] [truck]

(f) it's a [truck] [fire]

Both

Figure 3.2: Examples from the Winoground dataset for the swap-dependent linguistic tags *Object*, *Relation* and *Both* from left to right. They are additionally tagged with 1, 2 and 1 main predicates from left to right.

(a) the kid [with the magnifying glass] looks at them []

(b) the kid [] looks at them [with the magnifying glass]

Pragmatics

(c) the person with the ponytail [packs] stuff and other [buys] it

(d) the person with the ponytail [buys] stuff and other [packs] it

Series

(e) there are [three] people and [two] windows

(f) there are [two] people and [three] windows

Symbolic

Figure 3.3: Examples from the Winoground dataset for the visual tags *Pragmatics*, *Series* and *Symbolic* from left to right. They are additionally tagged with the *Relation* tag, and 1, 2, and 1 main predicate from left to right.

2.3 Experiments and Results

- Previous: zero-shot experiments with pre-trained models
  - CLIP, FLAVA, LXMERT, UniT, UNITER, VILLA, VinVL, ViLT, VisualBERT and ViLBERT
- Ours: zero-shot with pre-trained and fine-tuned models
  - OFA, BLIP, CLIP, FLAVA and ViLT
- Humans around 90%, models close to or below random chance (25%)
- Fine-tuning for retrieval and visual reasoning helps
- BLIP better than previous models, 10% in text score, 4% in image score and 7% in group score
  - Text score: whether the model selects the correct caption given an image
  - Image score: whether the model selects the correct image given a caption
  - Group score: combines text and image score, all combinations correct
- Still very far from humans, 40% gap in text scores, and 64% in image and group scores

2.3 Experiments and Results

- Swap-dependent linguistic tags:
  - Humans highest scores on Object, followed by Relation and then Both
  - Models opposite, highest scores on Both, shortest and least compositional
- Swap-independent linguistic tags:
  - Humans are better on 2 main predicates, longer and more complicated
  - Models opposite, best on 1 main predicate
- Visual tags:
  - Humans (~95%) and models (40%) good at Symbolic
  - Humans very bad at Pragmatics (50%), good at Series (95%)
  - Models very bad on Pragmatics and Series (0% in image and group scores)

3 VSR Zero-shot Experiments

- Visual Spatial Reasoning contains training and validation splits to train models
- VSR authors train and test three popular VLMs: VisualBERT, LXMERT, and ViLT
- We do zero-shot experiments with ViLT and BLIP fine-tuned on NLVR2

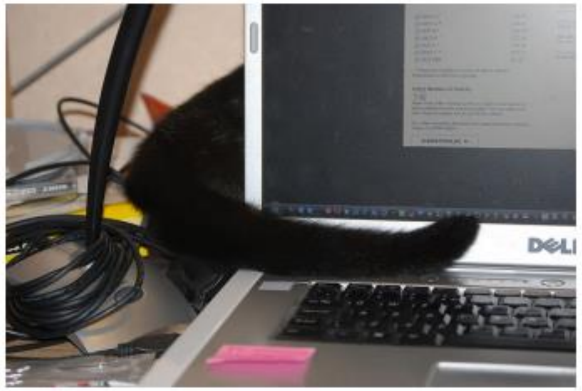
3.1 VSR Dataset

- Aims to test spatial grounding on natural COCO images
- Given an image and a caption, predict true or false
- Captions cover 65 spatial relations
- Grouped into 7 meta categories: Adjacency, Directional, Orientation, Projective, Proximity, Topological and Unallocated
- Random split: Split randomly into train/dev/test with a ratio of 70%/10%/20%.
- Zero-shot split: Train/dev/test sets have no overlapping concepts with a ratio of 50%/20%/30%.

(a) Caption: *The person is ahead of the cow.* Label: `True`.

(b) Caption: *The pizza is at the edge of the dining table.* Label: `True`.

Adjacency



(c) Caption: *The cat is behind the laptop.* Label: `True`.

(d) Caption: *The cat is behind the laptop.* Label: `False`.

Projective

(e) Caption: *The cat is inside the toilet.* Label: `False`.

(f) Caption: *The person is touching the hair drier.* Label: `True`.

Topological

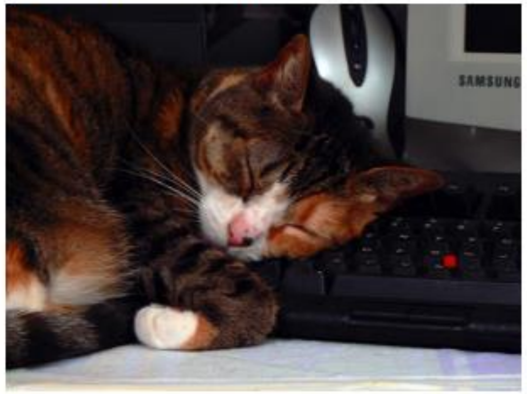
Figure 4.1: Examples from the VSR dataset for the relation meta categories *Adjacency*, *Projective* and *Topological* from left to right.

(a) Caption: *The potted plant is at the right side of the bench.* Label: `True`.

(b) Caption: *The cow is at the back of the car.* Label: `True`.

Adjacency

(c) Caption: *The bench is in front of the person.* Label: `True`.



(d) Caption: *The keyboard is below the cat.* Label: `True`.

Projective

(e) Caption: *The hair drier is facing away from the person.* Label: `False`.

(f) Caption: *The fire hydrant is facing away from the person.* Label: `True`.

Orientation

Figure 4.2: Examples from the VSR dataset for the relation meta categories *Adjacency*, *Projective* and *Orientation* from left to right.

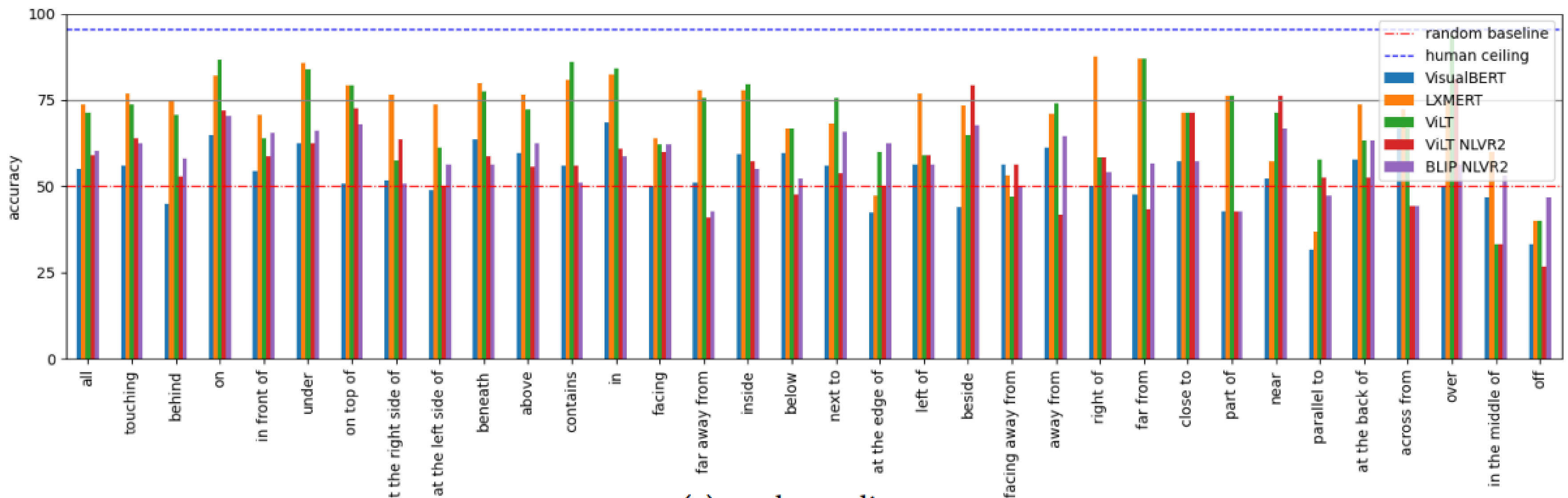
3.2 Experiments and Results

- VSR authors test three popular VLMs: VisualBERT, LXMERT and ViLT
- We also evaluate ViLT and BLIP fine-tuned on NLVR2
- Compared To Humans:
  - Random split: LXMERT and ViLT are the best over 70% accuracy. VisualBERT below 60%.
  - Zero-shot split: performance declines significantly. Best model 63%.
  - Compared to human performance, there is a more than 20% gap with the best models.
  - NLVR2 performance drop and dev/test difference is maintained.

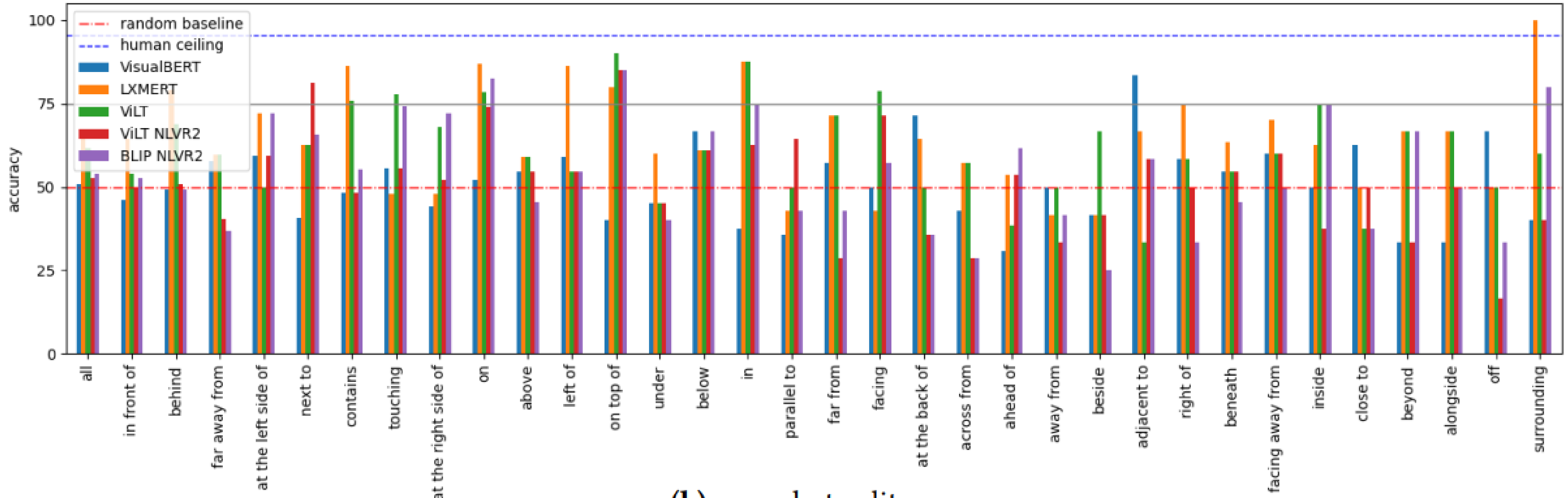
	random split		zero-shot split	
model↓	dev	test	dev	test
human	95.4			
VisualBERT	60.1	55.1	56.8	50.8
LXMERT	**73.3**	**73.9**	**70.3**	**65.5**
ViLT	72.7	71.2	66.0	61.6
ViLT NLVR2	57.9	59.1	56.4	52.8
BLIP NLVR2	60.9	60.1	57.9	53.9

3.2 Experiments and Results

- Some relations are harder, regardless of training examples
- Orientations and facing directions are very hard
- Left and right relations are difficult
- Examples can refer to either viewer’s or object’s reference frames
- Orientation worst on both splits, at chance level
- Proximity difficult on zero-shot, relative to concept and context

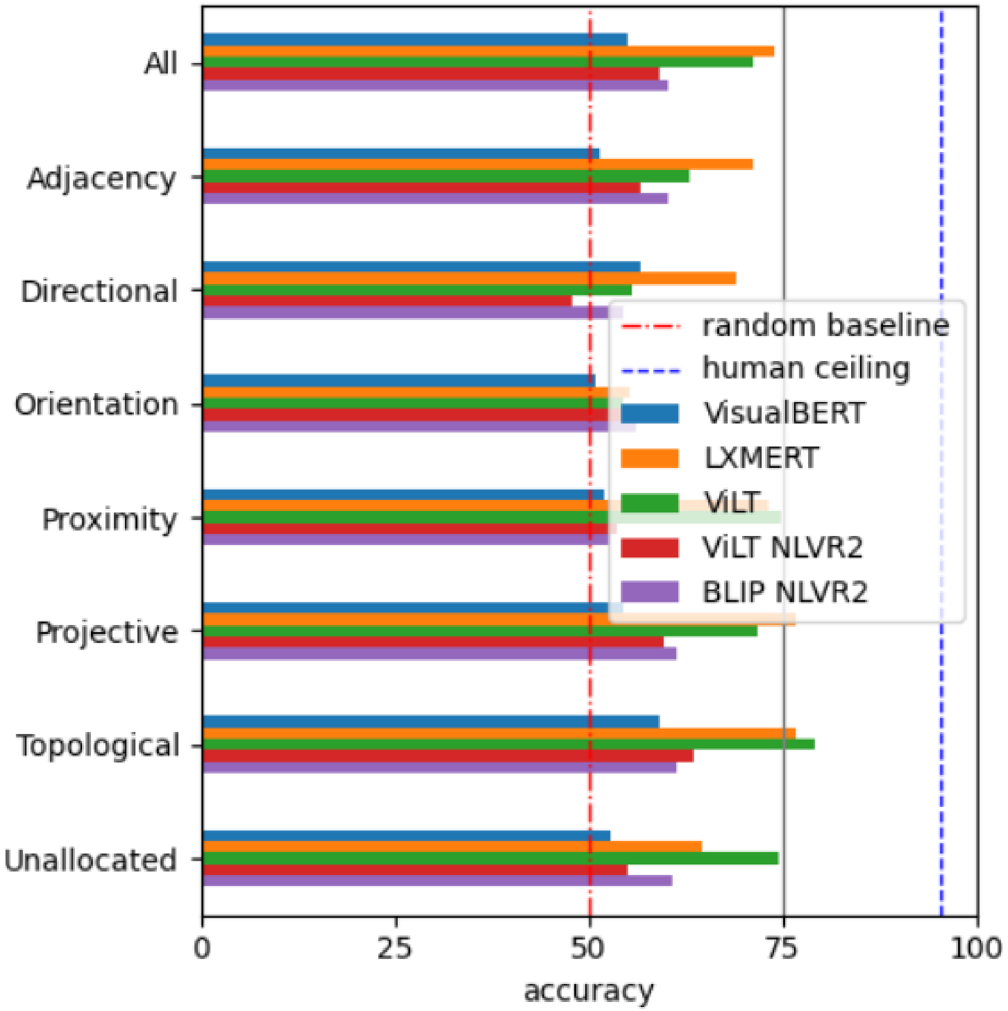


(a) random split

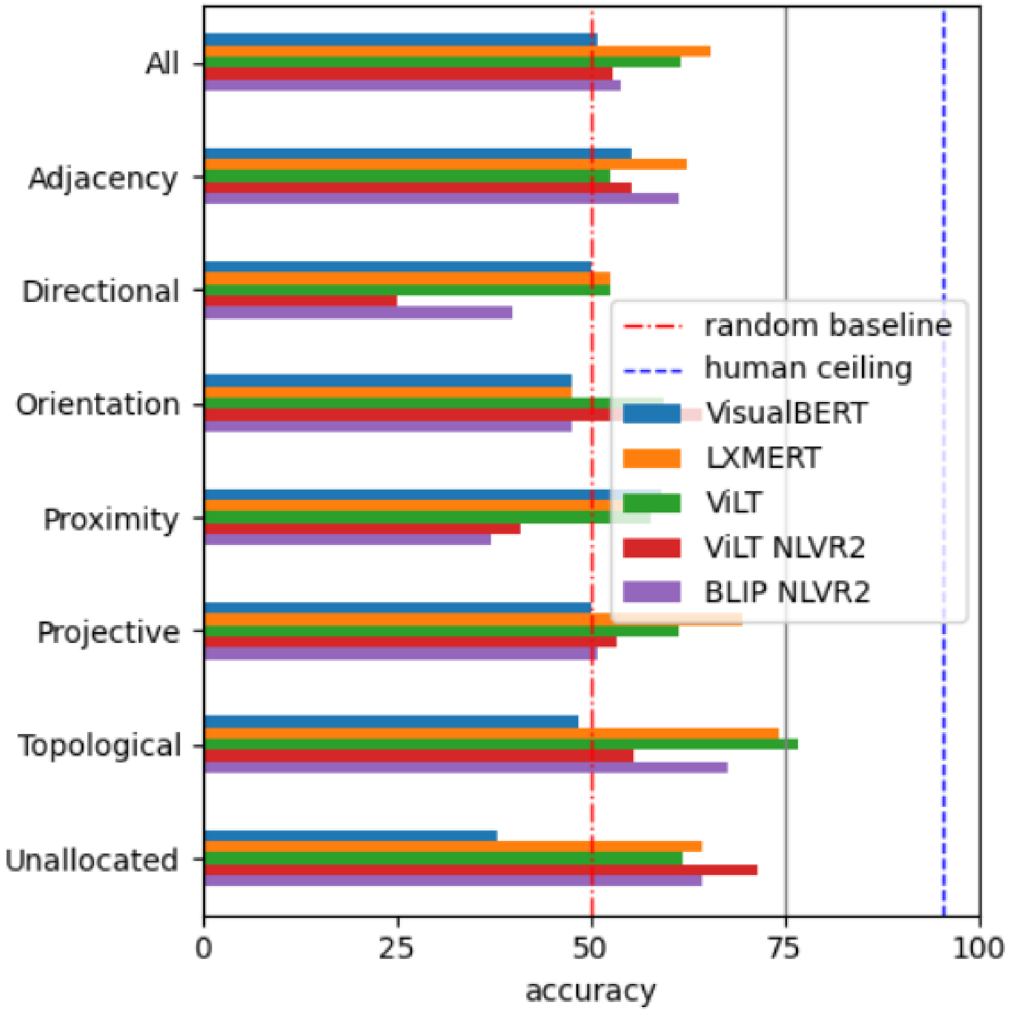


(b) zero-shot split

Figure 4.4: Our model performance by relation on the random (upper) and zero-shot (lower) split test sets. Relation order sorted by frequency (high to low from left to right). Only relations with more than 15 and 5 occurrences on the random and zero-shot tests respectively are shown.



(a) random split



(b) zero-shot split

Figure 4.6: Our model performance by meta categories of relations, on the random (left) and zero-shot (right) split test sets.

4 Synthetic Dataset Generation

- Winoground has no training data, annotating is time-consuming
- A solution could be to create a synthetic dataset for compositional reasoning
- Three options: Text-to-Image Generation, Image Captioning and Image Retrieval

4.1 Text-to-Image Generation

- We want to know if SD is good for synthetic dataset generation
- Stable Diffusion to generate images from Winogroud captions
- We also do a manual qualitative evaluation of the generated images
- 6 annotators in total and each one annotated 50 examples
- Total of 300 annotated examples and 600 images
- Conclusion is that SD is not good enough, most images are incorrect

	Caption 0	Caption 1	Both	None	All
Caption 0	65	48	12	175	300
Caption 1	46	65	13	176	300
All	111	113	25	351	600

(a) [some plants] surrounding [a light-bulb] ✓

(c) a [brown] dog is on a [white] couch ✗

(e) [circular] food on [heart-shaped] wood ✗

(b) [a lightbulb] surrounding [some plants] ✗

(d) a [white] dog is on a [brown] couch ✓

(f) [heart-shaped] food on [circular] wood ✗

Object *Relation* *Relation*

Figure 5.1: Stable Diffusion examples for the swap-dependent linguistic tags *Object*, *Relation* and *Relation* from left to right. They are additionally tagged with 1 main predicate. Correct examples are marked in green ✓ and incorrect ones in red ✗.

(a) there is [a mug] in [some grass] ✓
(b) there is [some grass] in [a mug] ✗
Object

(c) a person [sits] and a dog [stands] ✗
(d) a person [stands] and a dog [sits] ✓
Relation

(e) it's a [fire] [truck] ✓
(f) it's a [truck] [fire] ✓
Both

Figure 5.2: Stable Diffusion examples for the swap-dependent linguistic tags *Object*, *Relation* and *Both* from left to right. They are additionally tagged with 1, 2 and 1 main predicates from left to right. Correct examples are marked in green ✓ and incorrect ones in red ✗.

(a) the kid [with the magnifying glass] looks at them [] ✗

(c) the person with the ponytail [packs] stuff and other [buys] it ✗

(e) there are [three] people and [two] windows ✗

(b) the kid [] looks at them [with the magnifying glass] ✗

(d) the person with the ponytail [buys] stuff and other [packs] it ✗

(f) there are [two] people and [three] windows ✗

Pragmatics *Series* *Symbolic*

Figure 5.3: Stable Diffusion examples for the visual tags *Pragmatics*, *Series* and *Symbolic* from left to right. They are additionally tagged with the *Relation* tag, and 1, 2, and 1 main predicate from left to right. Correct examples are marked in green ✓ and incorrect ones in red ✗.

4.2 Image Captioning

- We evaluate image captioning for synthetic dataset creation
- OFA and BLIP models to generate captions for all Winoground images
- To evaluate, we calculated the BLEU score compared to real captions
- Captions are very different, but correct most of the times
- Generate accurate captions much faster than human annotation

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
OFA_{Tiny}	14.40	5.76	2.50	1.30
OFA_{Base}	16.68	7.12	3.26	1.58
OFA_{Medium}	16.28	6.47	2.84	1.39
OFA_{Large}	15.10	6.45	3.03	1.53
OFA_{Huge}	15.73	6.94	3.06	1.35
BLIP (ViT-B/16)	17.80	8.10	3.96	2.01
BLIP (ViT-L/16)	17.96	8.31	4.36	2.50

(a) a light bulb sitting on top of a pile of green leaves ✓

(b) a light bulb with a plant inside of it ✓

Object

(c) a black dog sitting on a couch in front of a christmas tree ✓

(d) a white dog sitting on top of a brown couch ✓

Relation

(e) a woman sprinkling herbs on a plate of food ✓

(f) a heart shaped pizza sitting on top of a cutting board ✓

Relation

Figure 5.5: Image Captioning examples from the Winoground dataset for the swap-dependent linguistic tags *Object*, *Relation* and *Relation* from left to right. They are additionally tagged with 1 main predicate. Correct examples are marked in green ✓ and incorrect ones in red ✗.

(a) a cup of coffee sitting on top of a lush green field ✓

(b) a cup with a plant in it sitting on a table ✓

Object

(c) a brown and white dog running in the sand ✓

(d) a man standing next to a dog in a kitchen ✓

Relation

(e) a red fire truck driving down a street ✓

(f) a car is on fire in a field ✓

Both

Figure 5.6: Image Captioning examples from the Winoground dataset for the swap-dependent linguistic tags *Object*, *Relation* and *Both* from left to right. They are additionally tagged with 1, 2 and 1 main predicates from left to right. Correct examples are marked in green ✓ and incorrect ones in red ✗.

(a) a man holding a magnifying glass next to a young boy ✗

(b) a man and a little girl sitting at a table ✗

Pragmatics

(c) a man and a woman wearing face masks in a store ✓

(d) a man and a woman in a grocery store ✓

Series

(e) a child's drawing of a house with a rainbow ✓

(f) a child's drawing of a house and a girl ✓

Symbolic

Figure 5.7: Image Captioning examples from the Winoground dataset for the visual tags *Pragmatics*, *Series* and *Symbolic* from left to right. They are additionally tagged with the *Relation* tag, and 1, 2, and 1 main predicate from left to right. Correct examples are marked in green ✓ and incorrect ones in red ✗.

4.3 Image Retrieval

- Retrieve images from LAION-5B using CLIP retrieval
- Retrieve similar images for each Winoground caption and image
- Use CLIP embedding similarity to retrieve similar images
- Remove non-aesthetic, duplicate and unsafe images
- Many images are wrong, would require filtering
- Could be combined with image captioning
- Low-effort dataset creation with minimal human work

(a) [some plants] surrounding [a light-bulb] ✓

(c) a [brown] dog is on a [white] couch ✗

(e) [circular] food on [heart-shaped] wood ✗

(b) [a lightbulb] surrounding [some plants] ✓

(d) a [white] dog is on a [brown] couch ✗

(f) [heart-shaped] food on [circular] wood ✓

Object | *Relation* | *Relation*

Figure 5.9: CLIP Retrieval examples for the swap-dependent linguistic tags *Object*, *Relation* and *Relation* from left to right. They are additionally tagged with 1 main predicate. Correct examples are marked in green ✓ and incorrect ones in red ✗.

(a) there is [a mug] in [some grass] ✓

(b) there is [some grass] in [a mug] ✓

Object

(c) a person [sits] and a dog [stands] ✗

(d) a person [stands] and a dog [sits] ✗

Relation

(e) it's a [fire] [truck] ✓

(f) it's a [truck] [fire] ✗

Both

Figure 5.10: CLIP Retrieval examples for the swap-dependent linguistic tags *Object*, *Relation* and *Both* from left to right. They are additionally tagged with 1, 2 and 1 main predicates from left to right. Correct examples are marked in green ✓ and incorrect ones in red ✗.

(a) the kid [with the magnifying glass] looks at them [] ✗

(b) the kid [] looks at them [with the magnifying glass] ✗

Pragmatics

(c) the person with the ponytail [packs] stuff and other [buys] it ✗

(d) the person with the ponytail [buys] stuff and other [packs] it ✗

Series

(e) there are [three] people and [two] windows ✗

(f) there are [two] people and [three] windows ✗

Symbolic

Figure 5.11: CLIP Retrieval examples for the visual tags *Pragmatics*, *Series* and *Symbolic* from left to right. They are additionally tagged with the *Relation* tag, and 1, 2, and 1 main predicate from left to right. Correct examples are marked in green ✓ and incorrect ones in red ✗.

5 Conclusions

- 5 main objectives accomplished
- Improve the state-of-the-art in compositional reasoning
  - Better than previous, still very far from humans
- Perform zero-shot experiments in spatial reasoning
  - Zero-shot performance drop, fine-tuning necessary
- Investigate text-to-image models for synthetic dataset creation
  - Not robust enough, a better approach needed
- Investigate image captioning for synthetic dataset creation
  - Different but good captions, can be combined with retrieval
- Investigate image retrieval for synthetic dataset creation
  - Many wrong images, some filtering needed

6 Future Work

- Four additional ideas for synthetic dataset generation
- Ideas for extending current datasets to be multilingual

6.1 Synthetic Dataset Generation

- Four additional ideas: explicit verbalization, text-to-image, image-to-image and image captioning and retrieval
- Use synthetic datasets to train VLMs in a self-supervised way
- Use multi-tasking and multi-sourcing
- Multi-tasking: learn more than one task simultaneously
- Multi-sourcing: combine different synthetic datasets

6.1.1 Explicit Verbalization

- Collect images with spatial relations (COCO)
- Use an object detector to identify the entities in the images
- Create verbalization templates by hand to generate captions

6.1.2 Text-to-Image Generation

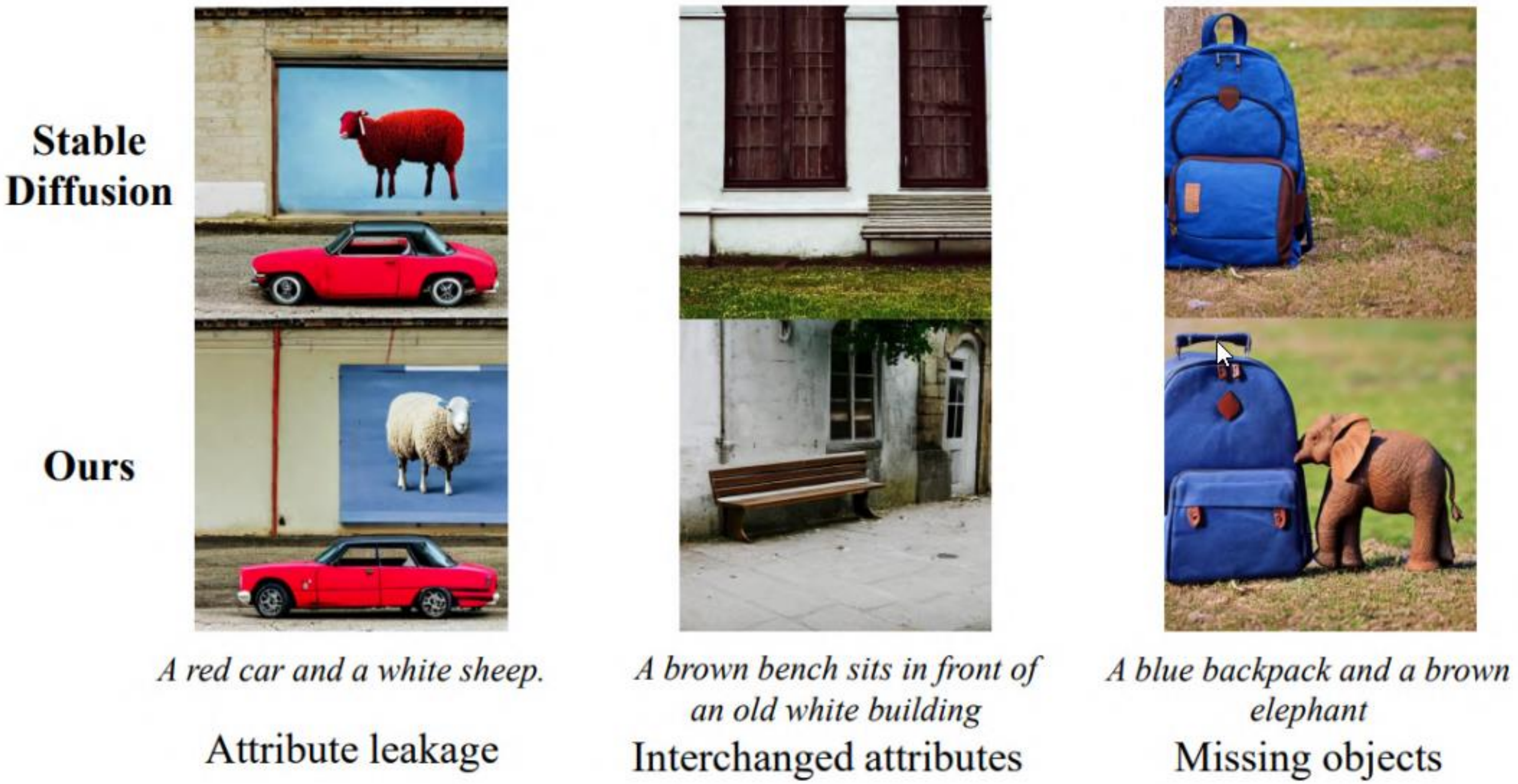


Figure 7.1: Three challenging phenomena in the compositional generation. Attribute leakage: The attribute of one object appears in another object. Interchanged attributes: the attributes of two or more objects are interchanged. Missing objects: one or more objects are missing.

6.1.3 Image-to-Image Generation

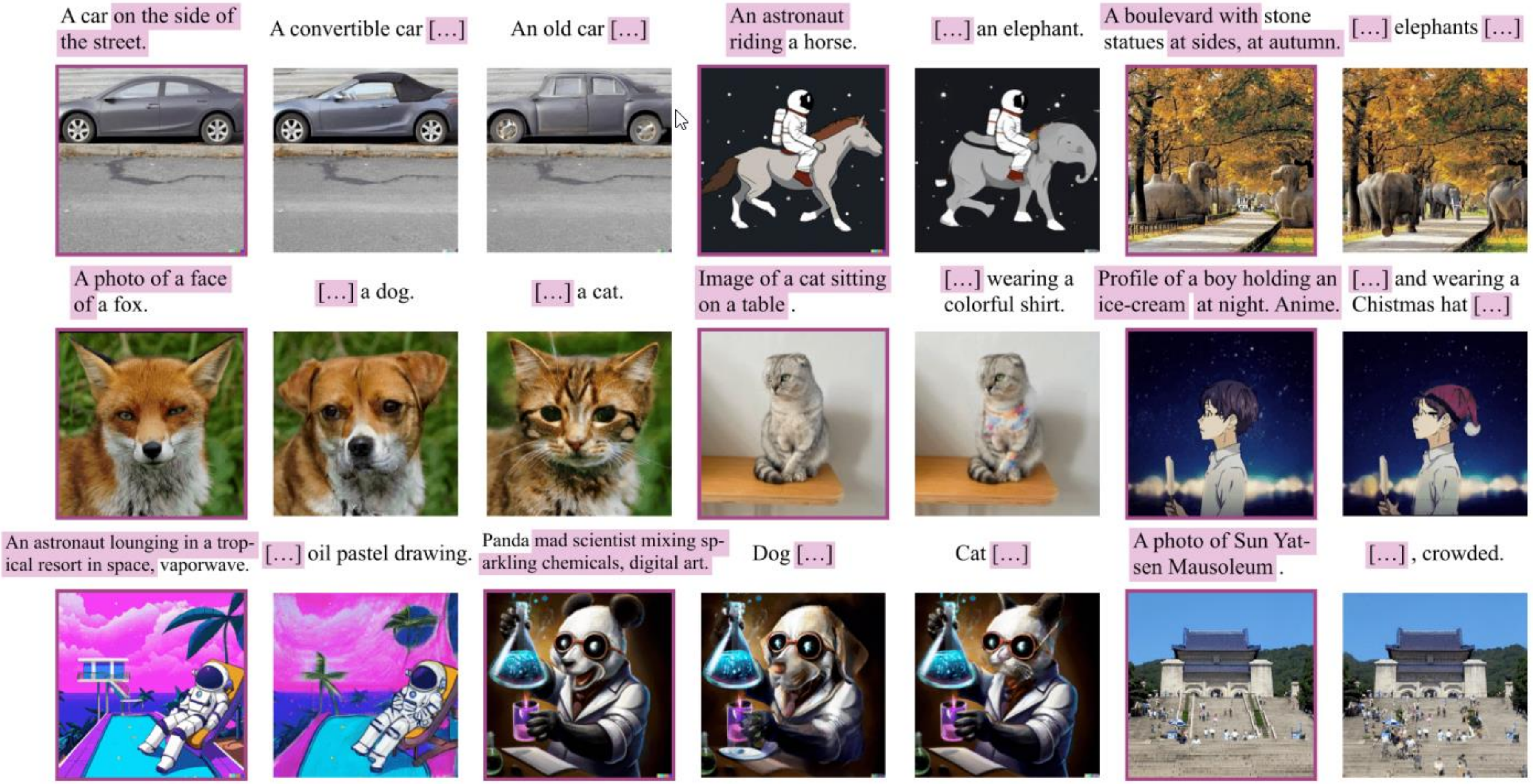


Figure 7.2: With CycleDiffusion text-to-image diffusion models can be used as zero-shot image-to-image editors. Source images are displayed with a purple margin and others are generated target images. CycleDiffusion achieves minimal editing that includes replacing objects, adding objects, changing image styles, and modifying attributes.

6.1.3 Image-to-Image Generation

Figure 7.3: Prompt-to-Prompt editing operations: tuning the level of influence of an adjective word (left), making a local modification in the image by replacing or adding a word (middle), or specifying a global modification (right).

6.1.3 Image-to-Image Generation

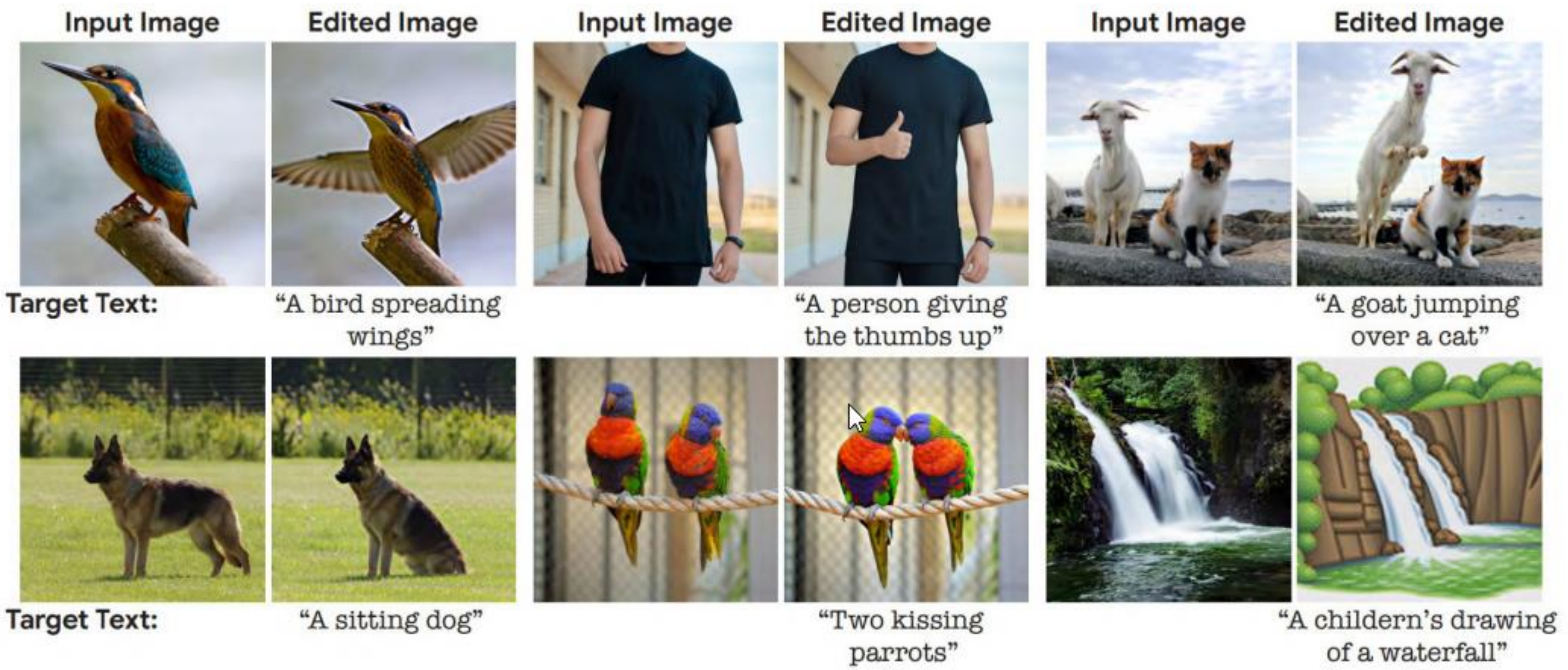


Figure 7.4: Imagic can perform various text-based semantic edits on a single real input image, including highly complex non-rigid changes such as posture changes and editing multiple objects. Here, we show pairs of input images and edited outputs with their respective target texts.

6.1.3 Image-to-Image Generation

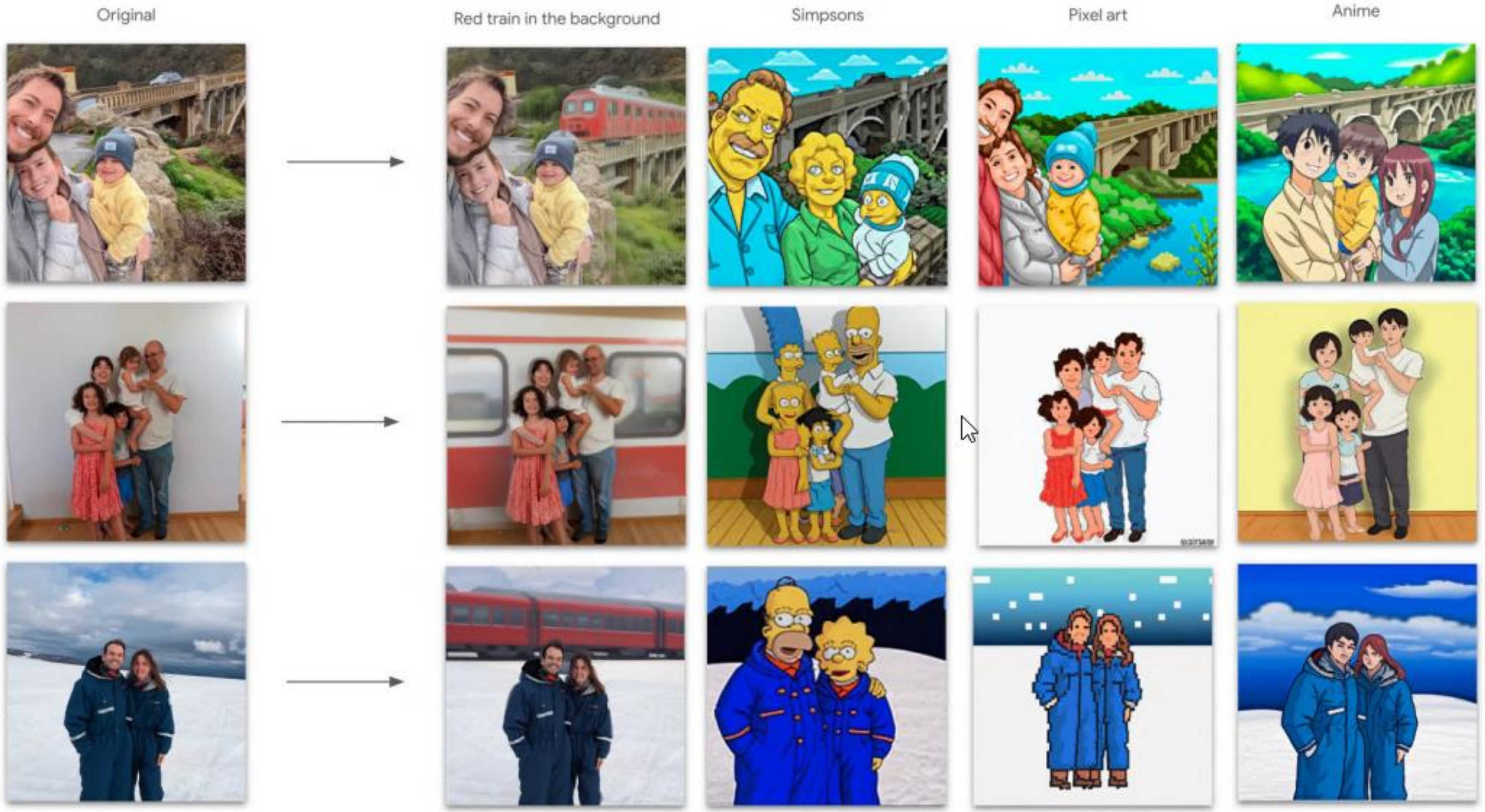


Figure 7.5: Samples showing UniTune's ability to maintain semantic details even across broad visual changes, and to place edits in a logical manner.

6.1.4 Image Captioning and Retrieval

- Retrieve images of interest from a huge dataset
- Use image captioning to create a dataset
- Similar example: LAION-COCO 600M
- BLIP to generate captions
- CLIP to rank captions
- T0 to correct grammar and punctuation errors

6.2 Multilingual Datasets

- Extend Winoground and VSR to more languages and cultures
- Winoground translation difficult
  - Both captions must contain the same words
  - Very difficult, impossible in some cases
- VSR translation seems easier
  - No word conditions
  - Different spatial relations across languages
  - Different word order in languages
- Multilingual pre-training datasets
  - LAION-2B-multi and LAION-1B-nolang
  - LAION-translated

Thank you!