

SUPPLEMENTARY MATERIALS FOR FUSING VISUAL AND TEXTUAL CUES FOR SEQUENTIAL IMAGE DIFFERENCE CAPTIONING

Anonymous authors

Paper under double-blind review

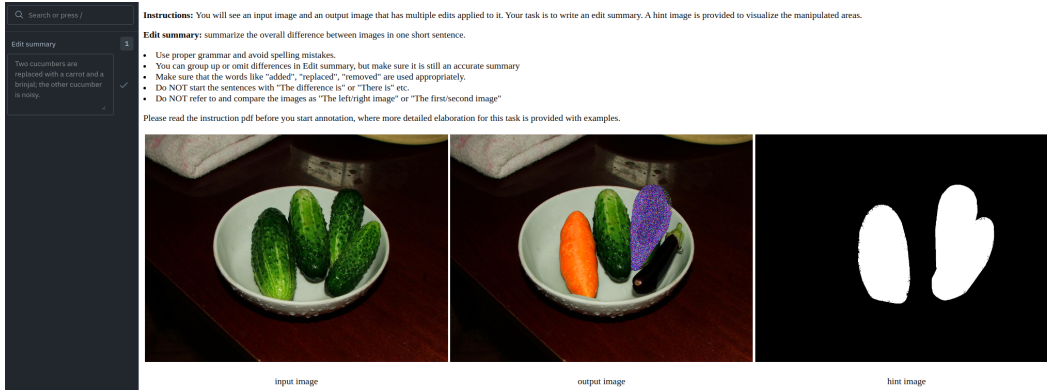


Figure 1: Example of a labeling task in Labelbox.

1 INFERENCE EXAMPLES

Fig. 2 shows inference examples on METS, CLEVR-Change, and Spot-the-Diff datasets, demonstrating the performance of FVTC-2 and FVTC-4 as well as GPT4V baseline.

2 LABELING PROCEDURE

We use the Labelbox platform ¹ to obtain the annotations for METS. The labelers are provided with a set of instructions and several examples, which can be found in an attached file `labeling_instructions.pdf`. An example of a task presented to the labelers is shown in Fig. 1. For each labeling task, the labelers are shown the original image and a manipulated image after either 5th, 10th or 15th edits. Additionally, a binary mask of the annotated images is shown to help guide the labeler’s attention. Overall, 17 labelers contributed 2,993 annotations with a mean of 176 and median of 185 annotations per person and an average of 5 min 27 sec per label.

2.1 SHORT SUMMARIES

We note that initially we have asked the labelers to provide a more detailed, exhaustive summary of the manipulations. However, we found that these captions were often too verbose and not very informative, similar to simply concatenating the machine generated list of manipulations.

3 FAILURE CASES

In Fig. 3 we show three examples of failure cases of FVTC. On the left, the model incorrectly uses a plural form. The model states that multiple oranges are replaced with multiple tennis

¹<https://www.labelbox.com/>

		
	Model	Output
	GPT4V (2):	The strawberry tart is replaced with an succulent plant, which is partially magenta, and the cake mould is now more visible. The second image has three pancakes replaced with a rubber duck, a toy car, and a denim clothing item; a shoe is added at the bottom edge of the image.
	GPT4V (4T):	Strawberries are replaced with noisy tomatoes and carrots, the cake is replaced with a coffee mug, and the tart is replaced with a carved and sliced watermelon. The potted plant that replaced the cake is given high contrast, striped overlay, and reduced encoding quality. The watermelon and coffee mug have altered saturation levels, and the carved watermelon's sharpness is decreased. The pancake is replaced with various objects such as a blurred and dark shoe, socks with reduced brightness, a toy car with enhanced contrast and dotted stripes, a computer disc with increased saturation, and a toothpaste tube with adjusted contrast. Additionally, bread is twice replaced with a toaster, which is manipulated with noise and severe blur. The overall image contrast is also increased severely.
	FVTC-2 (ours):	The slices of strawberries on the cake are slightly different, a green patch is added to a bowl. One pancake is replaced with a toy duck and one with a toy car.
	FVTC-4T (ours):	A strawberry pie is replaced with a potted plant, and one strawberry is covered with noise. A pancake is replaced with a toy car, others are replaced with a rubber duck, a sock and a woman's shoe.
	Target	The strawberry pie is replaced with an echeveria plant, and one strawberry is now noisy. Some pancakes are replaced with a car toy, a sock, a rubber duck, and a black shoe.
		
	GPT4V (2):	The drink in the martini glass has been altered with a green tint and a pattern resembling stripes and holiday lights. The overall color balance of the second image also seems slightly different, enhancing the green hue. Two doughnuts are manipulated in the second image: one is overlaid with a noisy, multicolored pattern and the other is now displaying a blue, partially transparent digital clock showing the time 9:15.
	GPT4V (4T):	The wine in the image is replaced with a champagne bottle which is then given solid green stripes, moderate brightness increase, and slight blur. The cocktail is replaced first by a coconut with reduced contrast and later by a seashell with decreased brightness and slightly decreased saturation. The juice is substituted with coffee adorned with dashed yellow stripes, followed by sparkling water with severely decreased sharpness. In a repeated manipulation, a champagne bottle replaces the wine again and undergoes severe sharpness reduction and moderate encoding quality decrease. The pastry is replaced multiple times with various objects such as a newspaper, toilet paper roll, flashlight, paintbrush, book, sticky notes, faucet, pencil, bookmark, and magazine. These objects undergo various manipulations like changes in brightness, contrast, sharpness, saturation, addition of random noise and overlay stripes, and some are blurred. The final image is significantly altered with decreased saturation and contrast
	FVTC-2 (ours):	A glass of drink is replaced with a glass of juice, and the other glass of drink is slightly blurred. A donut is replaced with a toilet paper roll, and a donut is covered with noise.
	FVTC-4T (ours):	A glass of drink is replaced with a glass of champagne and the second drink is covered with green patches. A donut is replaced with a flashlight, and another donut is covered with noise.
	Target	Drinks in the cocktail glasses are replaced with champagne, and one glass now has green lines on it. One bun is replaced with a torch, and a significant amount of noise has been applied to the other bun.
	CLEVR-Change	Spot-the-Diff
		
	FVTC-2 (ours):	The yellow shiny cube became gray. There is a person walking near the bottom of the picture on the left.
	Target	The large metallic cube changed to gray. In the after image there is an additional person not displayed in the first image.

Figure 2: Examples of model inference on METS (top 4), CLEVR-Change (bot-left), and Spot-the-Diff (bot-right) datasets. FVTC-2 and FVTC-4 denote the model trained on METS with 2 and 4 input images, respectively. "T" denotes the presence of auxiliary textual information.

balls, while the correct answer is that a single orange is replaced with a single tennis ball. In the middle, the model's caption is too vague and does not provide enough information about the manipulations. In the rightmost example, the model is miscounting the number



Figure 3: Failure cases of FVTC. Left: Incorrect plural form. Middle: Vague caption. Right: Miscounting and incorrect grammar.

of objects being manipulated and has incorrect grammar, stating "replaced with a blue ball and *another*", without specifying what the other object is.

4 TEXT PROMPTS

In Quotes 1,2 we provide the text prompts used for generation of the METS dataset. Quote 3 shows the prompt used for GPT4-V experiments with additional text inputs. The prompt used for GPT4-V without additional text inputs is similar, but omits the lists of manipulations and only includes the summary.

Quote 1: GPT3.5 text prompt for generation of replacement edits for METS.

Given a caption and an object choose what this object could be replaced with. The objects should be roughly the same size and shape. For object shape, refer to bbox which is given in (x1, x2, y1, y2) percentage coordinates. Creative and unexpected replacements are encouraged!

caption: A photo of a woman eating an apple tart
object: dessert
bbox: [0.14, 0.32, 0.71, 0.89]
replacement: chocolate ice-cream

caption: pink-animal-parade-crib-bedding
object: bed
bbox: [0.12, 0.63, 0.21, 0.84]
replacement: couch

caption: A-Line/Princess Scoop Neck Floor-Length Chiffon Evening Dress With Beading
object: person
bbox: [0.51, 0.65, 0.22, 0.75]
replacement: astronaut

Quote 2: GPT3.5 text prompt for generation of property change edits for METS.

Given a caption and an object choose what property of the object could be altered. This could be color, material. For object shape, refer to bbox which is given in (x1, x2, y1, y2) percentage coordinates. Creative and unexpected answers are encouraged!

caption: In this image there is a table, on that table there is a guitar.
object: table
bbox: [0.21, 0.78, 0.44, 0.68]
replacement: marble table

caption: pink-animal-parade-crib-bedding
object: bed
bbox: [0.12, 0.63, 0.21, 0.84]
replacement: japanese style bed

caption: A-Line/Princess Scoop Neck Floor-Length Chiffon Evening Dress With Beading
object: dress
bbox: [0.51, 0.64, 0.28, 0.62]
replacement: dress made out of aluminum foil

Quote 3: Text prompt for conditioning GPT4-V during inference

Given a list of manipulations, summarize the manipulations applied to make the last image from the first. Keep your answer to a few short sentences.

Example1:

List:

Object: Roller skates, replacement: background

Object: Roller skates, replacement: skateboards with LED lights

Object: Skateboard, replacement: background

Object: Skateboard, manipulation: encoding_quality, intensity: decreased slightly

Object: Skateboard, replacement: rollerblades

Object: rollerblades, manipulation: overlay_stripes, intensity: line width: 0.37, line color:

(170, 233, 137), line angle: 178, line density: 0.67, line type: dashed, line opacity: 0.62
 Object: rollerblades, manipulation: brightness, intensity: increased moderately
 Object: rollerblades, manipulation: encoding_quality, intensity: decreased moderately
 Object: Roller skates, replacement: background
 Object: Skateboard, replacement: rollerblades
 Object: rollerblades, manipulation: saturation, intensity: decreased moderately
 Summary: Shoes are now noisy; roller skates are replaced with a hover board and a patch.
 Example2:
 List:
 Object: Cheese, manipulation: brightness, intensity: decreased severely
 Object: Cheese, manipulation: contrast, intensity: decreased slightly
 Object: Cheese, replacement: pickle slice
 Object: pickle slice, manipulation: sharpness, intensity: increased moderately
 Object: pickle slice, manipulation: encoding_quality, intensity: decreased moderately
 Object: pickle slice, manipulation: saturation, intensity: decreased severely
 Object: pickle slice, manipulation: overlay_stripes, intensity: line width: 0.49, line color:
 (194, 157, 0), line angle: -85, line density: 0.089, line type: solid, line opacity: 0.73
 Object: Cheese, manipulation: contrast, intensity: increased moderately
 Object: Cheese, replacement: spicy cheese
 Object: Cheese, replacement: sushi
 Object: Cheese, replacement: crackers
 Object: Cheese, replacement: background
 Object: Cheese, replacement: crackers
 Object: crackers, replacement: crackers
 Object: global, manipulation: saturation, intensity: increased severely
 Summary: The cheese slices in the risotto bowls are replaced with a sunglass, chocolate pastry,
 an orange slice, a partially noisy pine cone, and a tennis ball.
 Your prompt:
 List: <list-of-edits>
 Summary: