Describing Differences in Image Sets with Natural Language

Supplementary Material

Acknowledgements

We thank all the reviewers for their constructive feedback. We thank James Zou, Weixin Liang, Jeff Z. HaoChen, Jen Weng, Zeyu Wang, Jackson Wang, Elaine Sui, Ruocheng Wang for providing valuable feedback to this project. We also thank Dan Klein for providing feedback on the abstract and intro as well as Esau Hutcherson and Yannis Siglidis for running preliminary experiments on VisDiffBench and the LaMem dataset. Lastly, we thank Alexei Efros for proposing several dozen applications, providing relevant related works, and for grudgingly acknowledging that the task of set difference captioning is "cool, even though it has language". This work was supported in part by the NSF CISE Expeditions Award (CCF-1730628). Trevor Darrell and Lisa Dunlap were supported by DoD and/or BAIR Industrial funds. Serena Yeung-Levy is a Chan Zuckerberg Biohub — San Francisco Investigator.

Reproducibility Statement

We provide code implementations of VisDiff at https://github.com/Understanding-Visual-Datasets/VisDiff. We also provide VisDiffBench at https://drive.google.com/file/d/1vghFd0rB5UTBaeR5rdxhJe3s7OOdRtkY. The implementations and datasets will enable researchers to reproduce all the experiments described in the paper as well as run their own analyses on additional datasets.

Ethics Statement

In this work, we introduce VisDiff, a novel method designed to discern subtle differences between two sets of images. VisDiff represents not just a technical advance in the analysis of image sets, but also serves as a useful tool to promote fairness, diversity, and scientific discovery in AI and data science. First, VisDiff has the potential to *uncover biases* in datasets. For instance, comparing image sets of workers from diverse demographic groups, such as men and women, can reveal and quantify career stereotypes associated with each group. This capability is pivotal in addressing and mitigating biases present in datasets. Furthermore, VisDiff holds substantial promise for *scientific discovery*. By comparing image sets in various scientific domains, such as cellular images from patients and healthy individuals, VisDiff can unveil novel insights into the disease impacts on cellular structures, thereby driving forward critical advancements in medical research. However, VisDiff is meant to be an *assistive tool* and should be applied with humans in the loop. The users are responsible for interpreting the results properly and avoiding misinformation. In summary, VisDiff emerges as a crucial tool for ethical AI considerations, fostering fairness and catalyzing scientific progress.

Table of Contents

In this supplementary material, we provide additional details of datasets, methods, results, and applications.

A. Supplementary Section 3

In this section, we provide additional details of Section 3 in the main paper.

A.1. Paired Sentences for VisDiffBench

VisDiffBench contains five subsets: PairedImageSets-Easy, PairedImageSets-Medium, PairedImageSets-Hard, ImageNetR, and ImageNet*. We provide all the paired sentences of PairedImageSets in Table 6. For ImageNetR, $\mathcal{D}_A$ is one of the "art", "cartoon", "deviantart", "embroidery", "graffiti", "graphic", "origami", "painting", "sculpture", "sketch", "sticker", "tattoo", "toy", "videogame", and $\mathcal{D}_B$ is "imagenet". For ImageNet*, $\mathcal{D}_A$ is one of the "in the forest", "green", "red", "pencil sketch", "oil painting", "orange", "on the rocks", "in bright sunlight", "person and a", "in the beach", "studio lighting", "in the water", "at dusk", "in the rain", "in the grass", "yellow", "blue", "and a flower", "on the road", "at night", "embroidery", "in the fog", "in the snow", and $\mathcal{D}_B$ is "base".

A.2. Examples for VisDiffBench

We provide 4 examples for PairedImageSets-Easy, PairedImageSets-Medium, PairedImageSets-Hard, respectively, in Figure 7 and Figure 16. For ImageNetR and ImageNet*, we refer readers to the original papers [14, 43].

Easy (50 Paired Sets)		Medium (50 Paired Sets)		Hard (50 Paired Sets)	
Set A	**Set B**	**Set A**	**Set B**	**Set A**	**Set B**
Dogs playing in a park	Cats playing in a park	SUVs on a road	Sedans on a road	Sunrise over Santorini, Greece	Sunset over Santorini, Greece
Children playing soccer	Children swimming in a pool	Wooden chairs in a room	Plastic chairs in a room	People practicing yoga in a mountainous setting	People meditating in a mountainous setting
Snow-covered mountains	Desert sand dunes	Golden retriever dogs playing	Labrador dogs playing	Fresh sushi with salmon topping	Fresh sushi with tuna topping
Butterflies on flowers	Bees on flowers	Green apples in a basket	Red apples in a basket	Lush vineyards in spring	Lush vineyards in early autumn
People shopping in a mall	People dining in a restaurant	Leather shoes on display	Canvas shoes on display	Men wearing Rolex watches	Men wearing Omega watches
Elephants in the savannah	Giraffes in the savannah	Freshwater fish in an aquarium	Saltwater fish in an aquarium	Cupcakes topped with buttercream	Cupcakes topped with fondant
Birds flying in the sky	Airplanes flying in the sky	Steel bridges over a river	Wooden bridges over a river	People playing chess outdoors	People playing checkers outdoors
Boats in a marina	Cars in a parking lot	Mountain bikes on a trail	Road bikes on a road	Hand-painted porcelain plates	Hand-painted ceramic plates
Tulips in a garden	Roses in a garden	Ceramic mugs on a shelf	Glass mugs on a shelf	Cyclists in a time-trial race	Cyclists in a mountain stage race
People skiing on a slope	People snowboarding on a slope	People playing electric guitars	People playing acoustic guitars	Gardens with Japanese cherry blossoms	Gardens with Japanese maples
Fish in an aquarium	Turtles in an aquarium	Laptop computers on a desk	Desktop computers on a desk	People wearing traditional Korean hanboks	People wearing traditional Japanese kimonos
Books on a shelf	Plants on a shelf	Hardcover books on a table	Paperback books on a table	Alpine lakes in summer	Alpine lakes in early spring
Grapes in a bowl	Apples in a bowl	Digital clocks on a wall	Analog clocks on a wall	Merlot wine in a glass	Cabernet Sauvignon wine in a glass
Motorcycles on a street	Bicycles on a street	Children playing with toy cars	Children playing with toy trains	Football players in defensive formation	Football players in offensive formation
Cows grazing in a field	Sheep grazing in a field	White roses in a vase	Pink roses in a vase	Classic novels from the 19th century	Modern novels from the 21st century
Babies in cribs	Babies in strollers	Electric stoves in a kitchen	Gas stoves in a kitchen	Orchestras playing Baroque music	Orchestras playing Classical music
Hot air balloons in the air	Kites in the air	Leather jackets on hangers	Denim jackets on hangers	Men in British army uniforms from WWI	Men in British army uniforms from WWII
Penguins in the snow	Seals in the snow	People eating with chopsticks	People eating with forks	Sculptures from the Renaissance era	Sculptures from the Hellenistic era
Lions in a jungle	Monkeys in a jungle	Pearl necklaces on display	Gold necklaces on display	People preparing macarons	People preparing meringues
Watches on a display	Rings on a display	Mushrooms in a forest	Ferns in a forest	Female ballet dancers in pointe shoes	Female ballet dancers in ballet slippers
Pizzas in a box	Donuts in a box	Stainless steel kettles in a store	Plastic kettles in a store	Dishes from Northern Italian cuisine	Dishes from Southern Italian cuisine
Bricks on a wall	Tiles on a wall	Porcelain vases on a shelf	Metal vases on a shelf	Classic rock bands performing	Alternative rock bands performing
Pianos in a room	Guitars in a room	Vintage cars on a road	Modern cars on a road	Historical films set in Medieval Europe	Historical films set in Ancient Rome
Trains on tracks	Buses on roads	Handmade quilts on a bed	Factory-made blankets on a bed	Bonsai trees shaped in cascade style	Bonsai trees shaped in informal upright style
Pots on a stove	Plates on a table	Shiny silk dresses on mannequins	Matte cotton dresses on mannequins	Lace wedding dresses	Satin wedding dresses
Stars in the night sky	Clouds in the day sky	Mechanical pencils on a desk	Ballpoint pens on a desk	Birds with iridescent plumage	Birds with matte plumage
Sunflowers in a field	Wheat in a field	Ginger cats lying down	Tabby cats lying down	Women wearing matte lipstick	Women wearing glossy lipstick
Dolls on a shelf	Teddy bears on a shelf	People riding racing horses	People riding dressage horses	Cities with Gothic architecture	Cities with Modernist architecture
Pine trees in a forest	Oak trees in a forest	Steel water bottles on a table	Glass water bottles on a table	Poems written in free verse	Poems written in sonnet form
Men playing basketball	Women playing volleyball	Men wearing leather gloves	Men wearing wool gloves	Acoustic guitars being played	Classical guitars being played
Ice cream in a cone	Juice in a glass	Rubber ducks in a tub	Plastic boats in a tub	Books with hardcover binding	Books with leather-bound covers
Dancers on a stage	Singers on a stage	Porcelain tea cups on a tray	Glass tea cups on a tray	Portraits painted in cubist style	Portraits painted in impressionist style
Rainbows in the sky	Lightning in the sky	Sparrows on a tree	Canaries on a tree	Residential buildings in Art Deco style	Residential buildings in Brutalist style
Towers in a city	Houses in a suburb	Shiny metallic cars	Matte finish cars	Male professional swimmers in freestyle race	Male professional swimmers in butterfly race
Frogs by a pond	Ducks by a pond	Stuffed teddy bears on a bed	Stuffed bunny rabbits on a bed	Basketball players attempting free throws	Basketball players attempting slam dunks
Football players on a field	Rugby players on a field	Round dinner plates on a table	Square dinner plates on a table	Cakes decorated with marzipan	Cakes decorated with buttercream roses
Pillows on a bed	Blankets on a bed	Butter on a slice of bread	Jam on a slice of bread	People practicing the Sun Salutation in yoga	People practicing the Tree Pose in yoga
Deer in a forest	Rabbits in a forest	Bengal cat in sitting posture	Siamese cat in sitting posture	Men wearing suits	Men wearing tuxedos
Tea in a cup	Coffee in a cup	Violinists playing in a quartet	Cellists playing in a quartet	Butterflies with spotted wings	Butterflies with striped wings
Children on a slide	Children on a swing	Gothic cathedrals in Europe	Baroque churches in Europe	Oak trees in summer	Oak trees in autumn
Kangaroos in a desert	Camels in a desert	People dancing tango	People dancing waltz	Tennis shoes on a rack	Running shoes on a rack
Tomatoes in a basket	Eggs in a basket	Abstract oil paintings with warm colors	Abstract oil paintings with cool colors	People playing classical violin	People playing fiddle
People in an elevator	People on an escalator	Candies made from dark chocolate	Candies made from milk chocolate	Men wearing fedoras	Men wearing baseball caps
Sandcastles on a beach	Umbrellas on a beach	Rivers in tropical rainforests	Rivers in alpine meadows	Passenger planes in the sky	Cargo planes in the sky
Mice in a barn	Horses in a barn	Cars from the 1960s	Cars from the 1980s	Women wearing ankle boots	Women wearing knee-high boots
Chocolates in a box	Candies in a jar	Seascapes during a storm	Seascapes during a calm day	Diesel trucks on a highway	Electric trucks on a highway
Zebra crossings on a street	Traffic lights on a street	Fruits arranged in a still life setting	Flowers arranged in a still life setting	Children reading comic books	Children reading fairy tales
Bridges over a river	Boats on a river	Dishes from Thai cuisine	Dishes from Vietnamese cuisine	Men wearing round glasses	Men wearing square glasses
Oranges on a tree	Bird nests on a tree	Wild horses in American plains	Wild zebras in African savannahs	Vinyl records in a store	CDs in a store
Lanterns in a festival	Fireworks in a festival	Classic movies in black and white	Classic movies in Technicolor	Bonsai trees in pots	Cacti in pots

Table 6. Paired sentences for PairedImageSets. Easy, medium, and hard examples are shown in the left, middle, and right.

A.3. Prompts for VisDiffBench Generation

We provide the GPT-4 prompt we used to generate paired sentences for PairedImageSets in Figure 8 (top).

A.4. Prompts for VisDiffBench Evaluation

We provide the GPT-4 prompt we used to evaluate the generated difference description against the ground-truth difference description in Figure 8 (bottom).

A.5. Human-generated Differences for VisDiffBench

To increase the quality of the dataset, we have collected human-generated differences between the sets in VisDiffBench. We have conducted two types of human annotations: **(1)** propose the differences by humans; **(2)** validate the differences by humans. Averaged across **3** annotations for each of **187** sets, we find that annotators agreed that **96%** of our labels are correct differences, **93%** are the best description to differentiate the set, and **76%** are the same as a difference the annotator has written. The last statistic is indicative of human performance on this challenging

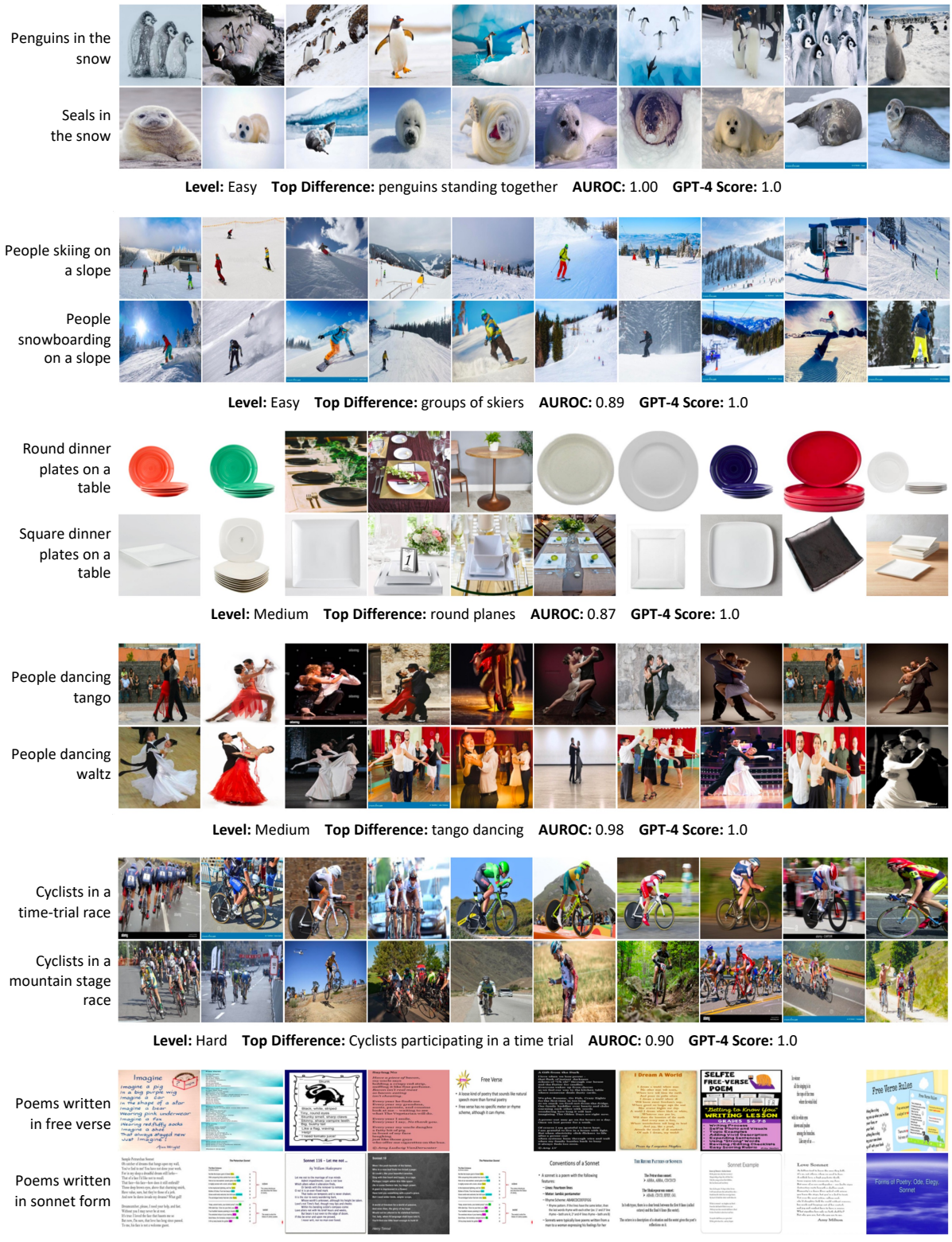


Figure 7. PairedImageSets examples where VisDiff succeeds. We show the ground-truth difference, top difference predicted by VisDiff, AUROC score output by the ranker, and evaluation of the predicted difference by GPT-4.

Metric	B4	R1	RL	BS	V1.5	G3.5	G4
Pearson	0.140	0.492	0.497	0.272	0.594	0.623	**0.800**

Table 7. **Correlation of automated metric with humans.** Model-free metrics include B4 (BLEU-4), R1 (ROUGE-1), RL (ROUGE-L). Model-based metrics include BS (BERTScore), V1.5 (Vicuna-1.5-13B), G3.5 (gpt-3.5-turbo-0613), G4 (gpt-4-0613).

task. Since this task has some difficult cases, for instance, when set A is "cities with Gothic architecture" and set B is "cities with Modernist architecture", we see models outperform humans on some cases.

In the first part of annotation, annotators are given the link to the images from $\mathcal{D}_{\mathcal{A}}$ and $\mathcal{D}_{\mathcal{B}}$ and asked to propose up to 5 differences (usually 1-2). In the second part, the annotators are given our VisDiffBench ground-truth descriptions and asked (1) is the provided difference correct (2) would you consider this the best description of the difference between the sets (3) is this consistent with any of your descriptions and (4) which description is it most consistent with. We gave each annotator a tutorial on the task with 3 examples, checking their first few descriptions were in the correct format. In the end we collected 3 annotations per image set in VisDiffBench. The inter-annotator agreement is 93%, 87%, 75% for questions 1-3. We have released these human labels along with our original labels in our code base.

A.6. Other VisDiffBench Evaluation Metrics

We chose GPT-4 as our evaluation metric because evaluating the proposed difference requires a human-level understanding of the semantics in a short description. Table 7 reports the correlation between common captioning metrics and human annotators in VisDiffBench, which shows that the GPT-4 evaluation has much higher consensus with humans and is the only reliable metric. However, due to the limitation of GPT-4 being closed-source and constantly changing, we highlighted the exact GPT version we used (*gpt-4-0613*) and released the outputs of our experiments to maximize reproducibility.

B. Supplementary Section 4

In this section, we provide additional details of Section 4 in the main paper.

B.1. Details for Proposer

We ran each proposer for 3 rounds. For each round, we sample 20 examples per set and generate 10 hypotheses.

Image-based Proposer. We provide an example input of the gridded image in Figure 9. We feed the image and the prompt shown in Figure 10 (middle) to LLaVA-1.5 to generate 10 hypotheses.

Feature-based Proposer. To generate 10 hypotheses, we sample BLIP-2 10 times using top-p sampling given the subtracted embedding.

Caption-based Proposer. We generate captions for each image using BLIP-2 with the prompt "Describe this image in detail.". We apply a simple filtering of the captions, which removes any special tokens and captions simply repeating words. We feed the filtered captions and the prompt shown in Figure 10 (top) to GPT-4 to generate 10 hypotheses.

Metric	Easy		Medium		Hard		IN-R/*	
	A1	A5	A1	A5	A1	A5	A1	A5
AUROC	**0.88**	**0.99**	**0.75**	**0.86**	**0.61**	**0.80**	0.78	**0.96**
p-value	0.83	**0.99**	0.74	**0.86**	0.58	0.77	**0.81**	0.95
diff. in means	0.83	0.98	0.69	0.84	0.60	0.76	0.76	0.92

Table 8. **VisDiffBench results using different ranking metrics based on CLIP similarity scores.** We use the caption-based proposer. A1 & A5 are Acc@1 & Acc@5.

B.2. Details for Ranker

Image-based Ranker. Given each hypothesis, we prompt LLaVA-1.5 with the image and the prompt "Does this image contain {hypothesis}?".

Caption-based Ranker. Given each hypothesis, we prompt Vicuna-1.5 with the image caption and hypothesis using the prompt shown in Figure 10 (bottom).

Feature-based Ranker. We use the OpenCLIP model ViT-bigG-14 trained on laion2b_s39b_b160k.

B.3. Different Ranking Metrics

Table 8 shows the results of several different ranking metrics using the CLIP similarity scores on VisDiffBench. We see that AUROC produces the most consistent highest performing results, but other metrics such as p-value and the difference in means also produce promising results.

C. Supplementary Section 5

In this section, we provide additional details of Section 5 in the main paper. We ablate various design choices of VisDiff.

VisDiffBench Generation Prompt

I'm working on a project about explaining image distributional difference using natural language. The inputs are image set A and image set B, the output is a natural language description of the most different features.

However, I don't have datasets to evaluate the system. I'm going to crawl Google to collect images giving a sentence. Can you think about 50 paired sentences showing:
- Easy level (compare super class, e.g., "Dogs playing in a park" vs "Cats playing in a park", "Children playing soccer" vs "Children swimming in a pool)
- Medium level difference (compare fine-grained class, e.g., "SUVs on a road" vs "Sedans on a road", "Wooden chairs in a room" vs "Plastic chairs in a room"),
- Difficult level (compare difficult, non-trivial differences, e.g., "sunrise over Santorini, Greece" vs "Sunset over Santorini, Greece", "Bengal cat in sitting posture" vs "Siamese cat in sitting posture")?

For each level, only includes one difference between two set (e.g., there are two differences between "Mountains in winter" vs "Beaches in summer", both scene and season, do not include this). Give 50 outputs in jsonl format ```{"set1": str, "set2": str, "difference": str}```. Let us start with 50 easy examples.

You did a great job! Let's do 50 medium level

Great! Let's do 50 difficult examples

VisDiffBench Evaluation Prompt

I am a machine learning researcher summarizing differences in groups of images. The goal is to find a concept that is more true for Group A than Group B.

Given a description of Group A and Group B, output whether a given prediction aligns with the description of Group A. Answer with a 2 (fully aligned), 1 (somewhat aligned), or 0 (not aligned). a score of 1 should be given if the prediction is more true for A than B, but is a superset or a subset of the most correct difference.

For example, if Group A is "images of dogs in the snow" and Group B is "images of dogs next to cats":
- predictions like "dogs in the snow" or "dogs in winter time" should be given a 2
- predictions like "golden retrievers on a ski slope" or "animals in the snow" should be given a 1

Here is the descriptions
Group A: People riding racing horses and Group B: People riding dressage horses. Prediction: Horse racing events

Again, output either a 2, 1, or 0. Response:

Figure 8. Prompt used to generate paired sentences for VisDiffBench (top) and evaluate VisDiffBench (bottom). Input-dependent texts are colored in blue.

Proposer	Ranker	PIS-Easy		PIS-Medium		PIS-Hard		ImageNet-R/*	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
GPT-4 on BLIP-2 Captions	CLIP	**0.88**	**0.99**	**0.75**	**0.86**	**0.61**	**0.80**	0.78	**0.96**
GPT-4 on LLaVA-1.5 Captions	CLIP	**0.89**	**0.98**	**0.73**	**0.85**	0.51	0.70	**0.84**	0.93
GPT-3.5 on BLIP-2 Captions	CLIP	0.81	0.95	0.67	**0.87**	**0.60**	0.76	**0.85**	**0.96**

Table 9. Results on VisDiffBench with different captions and language models. We bold any numbers within 0.02.

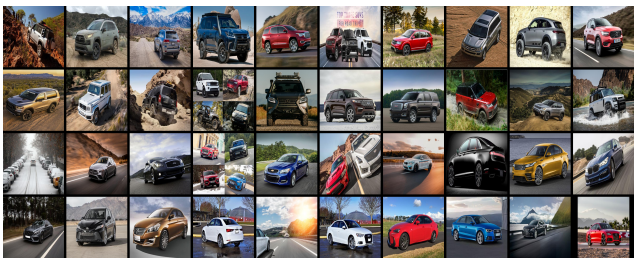


Figure 9. Example input to the image-based proposer. We arrange 20+20 input images into a single 4-row, 10-column gridded image.

C.1. Caption Styles

Given that our leading proposer is caption-based, it naturally raises the question of how captions derived from vision language models influence performance. We conducted a comparative analysis of captions generated by two state-of-the-art vision language models: BLIP-2 and LLaVA-1.5. Notably, compared to BLIP-2, LLaVA-1.5 has been instruction-tuned and can produce captions that are much longer with detailed information. The average caption length for LLaVA is around 391 characters compared to BLIP-2's 41 characters. As shown in Table 9, despite the clear disparity between these two captions, the algorithm achieves similar performances. This suggests that language models possess a robust inductive reasoning ability that allows them to discern the most notable differences in language. BLIP-2's captions, being marginally superior, could be attributed to their shortness and conciseness.

C.2. Language Models

We compared GPT-4 with GPT-3.5 in Table 9 to assess how different language models affect the caption-based proposer. While both models achieve strong performances on VisDiffBench, GPT-4 outperforms GPT-3.5 in most cases, demonstrating that the stronger reasoning capability of language models is important to accomplish the set difference captioning task.

Caption-based Proposer Prompt

The following are the result of captioning two groups of images:

Group A: a group of jockeys and horses are racing on a green field
Group A: a cowboy riding a bucking horse at a rodeo
...

Group B: a person is riding a black horse in an arena
Group B: person riding a horse in an equestrian competition at the london 2012 olympics
...

I am a machine learning researcher trying to figure out the major differences between these two groups so I can better understand my data.

Come up with 10 distinct concepts that are more likely to be true for Group A compared to Group B. Please write a list of captions (separated by bullet points "*"). For example:
* "a dog next to a horse"
* "a car in the rain"
* "low quality"
* "cars from a side view"
* "people in a intricate dress"
* "a joyful atmosphere"

Do not talk about the caption, e.g., "caption with one word" and do not list more than one concept. The hypothesis should be a caption, so hypotheses like "more of ...", "presence of ...", "images with ..." are incorrect. Also do not enumerate possibilities within parentheses. Here are examples of bad outputs and their corrections:
* INCORRECT: "various nature environments like lakes, forests, and mountains" CORRECTED: "nature"
* INCORRECT: "images of household object (e.g. bowl, vacuum, lamp)" CORRECTED: "household objects"
* INCORRECT: "Presence of baby animals" CORRECTED: "baby animals"
* INCORRECT: "Different types of vehicles including cars, trucks, boats, and RVs" CORRECTED: "vehicles"
* INCORRECT: "Images involving interaction between humans and animals" CORRECTED: "interaction between humans and animals"
* INCORRECT: "More realistic images" CORRECTED: "realistic images"
* INCORRECT: "Insects (cockroach, dragonfly, grasshopper)" CORRECTED: "insects"

Again, I want to figure out what kind of distribution shift are there. List properties that hold more often for the images (not captions) in group A compared to group B. Answer with a list (separated by bullet points "*"). Your response:

Image-based Proposer Prompt

This image contains two groups of images. 20 images from Group A are shown in the first two rows, while 20 images from Group B are shown in the last two rows.

I am a machine learning researcher trying to figure out the major differences between these two groups so I can better understand my data.

Come up with 10 distinct concepts that are more likely to be true for Group A compared to Group B. Please write a list of captions (separated by bullet points "*"). For example:
* "a dog next to a horse"
* "a car in the rain"
... (same as caption-based proposer prompt)
* INCORRECT: "Insects (cockroach, dragonfly, grasshopper)" CORRECTED: "insects"

Again, I want to figure out what kind of distribution shift are there. List properties that hold more often for the images in group A compared to group B. Answer with a list (separated by bullet points "*"). Your response:

Caption-based Ranker Prompt

Given a caption and a concept, respond with yes or no.
Here are 5 examples for the concept "spider and a flower":
INPUT: a spider sitting on top of a purple flower
OUTPUT: yes
INPUT: a yellow and black spider with a web in the background
OUTPUT: no
INPUT: a arachnid with a white flower
OUTPUT: yes
INPUT: a spider is walking on the ground in the grass
OUTPUT: no
INPUT: two yellow and black spiders
OUTPUT: no

Here are 6 examples for the concept "an ipod in the forest":
INPUT: a smartphone in the forest
OUTPUT: yes
INPUT: a white apple ipad sitting on top of a wooden table
OUTPUT: no
INPUT: an ipod near some trees
OUTPUT: yes
INPUT: a smartphone with apps
OUTPUT: no
INPUT: a pink mp3 player sitting on top of a book
OUTPUT: no
INPUT: an ipod sitting on a white surface
OUTPUT: no

Given the caption "mario and luigi are playing tennis on a white background" and the concept "references to pop culture", respond with either the word yes or no ONLY.
OUTPUT:

Figure 10. Prompt for caption-based proposer (top), image-based proposer (middle), and caption-based ranker (bottom). Input-dependent texts are colored in blue.

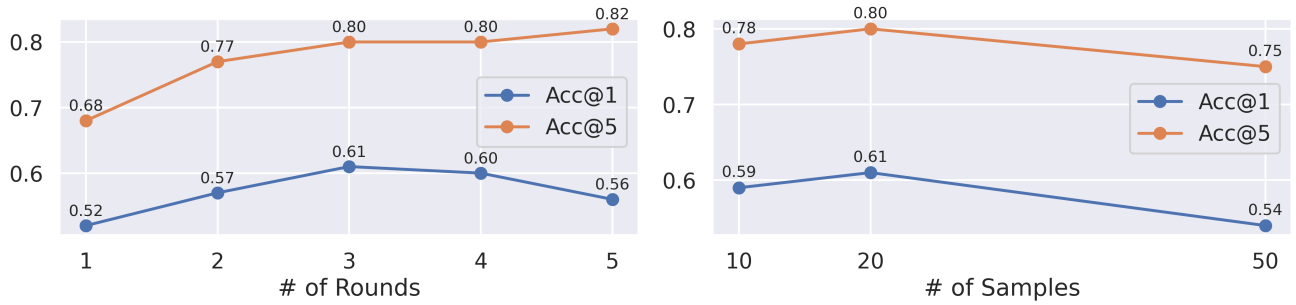


Figure 11. Analysis of the number of rounds (left) and number of samples (right) for the proposer on 50 PairedImageSets-Hard sets. 3 rounds and 20 samples are the best in terms of performance and efficiency.

C.3. Sampling Rounds

The proposer's generated differences rely on the random samples drawn from each image set; thus, extensive sampling is paramount to capture all the differences. Our ablation studies presented in Figure 11 (left), conducted on the PairedImageSets hard subset, suggest that an increase in sampling iterations typically correlates with enhanced performance. However, a point of diminishing returns is observed beyond three rounds of proposals. In this paper, we standardize the experiments based on three rounds of proposal generation.

C.4. Number of Sampled Examples

Inputting more samples from $\mathcal{D}_A$ and $\mathcal{D}_B$ into the proposer may not be advantageous, as a long context could result in information getting lost in the middle [26, 40]. Results shown in Figure 11 (right) reflect this, as inputting more captions to the large language models sees performance benefits up to 20 captions, at which point performance degrades.

C.5. Necessity of Ranker

Since the proposer may already generate and sort the most salient difference descriptions based on the sampled images, we conduct ablations to understand whether the ranker is necessary. We observe that, on PairedImageSets hard subset, VisDiff achieves 0.54 Acc@1 and 0.68 Acc@5 without ranker, which is much lower than the 0.61 Acc@1 and 0.80 Acc@5 with ranker, demonstrating the necessity of the ranker.

D. Supplementary Section 6

In this section, we provide additional details of Section 6 in the main paper.

D.1. Comparing ImageNetV2 with ImageNet

Per-class visualizations. Along with the "Dinner Table" example shown in Figure 1, we provide other per-class differences with the highest difference scores in Figure 12. These examples clearly reveal salient differences between ImageNetV2 and ImageNet. Moreover, we observe time differences between these two datasets, as ImageNetV2 images contain Twitter and WhatsApp in the "Hand-held Computer" class and London 2012 Oylmpics in the "Horizontal Bar" class.

ImageNetV2 metadata analysis. To get more precise statistics on when the ImageNetV2 images were collected, we analyzed the metadata of each image, which reports the minimum and maximum upload date of that image to Flickr. We find that 72% images were uploaded between 2012 and 2013, and 28% were uploaded between 2013 and 2014. This is different from ImageNet images that were all collected on or before 2010.

D.2. Comparing Behaviors of CLIP and ResNet

Top Differences and Per-class visualizations. We provide per-class differences where CLIP outperforms ResNet most in Figure 13. These examples clearly reveal salient differences between CLIP and ResNet, such as CLIP's robustness to label within images, object angle, and presence of people.

D.3. Finding Failure Modes of ResNet

Model details. We use the PyTorch pre-trained ResNet-50 and ResNet-101 models and the Huggingface "facebook/detr-resnet-50" object detector.

Top differences. The top 5 difference descriptions from VisDiff were "humanized object items", "people interacting with objects", "electronics and appliances", "objects or people in a marketplace setting", and "household objects in unusual placement".

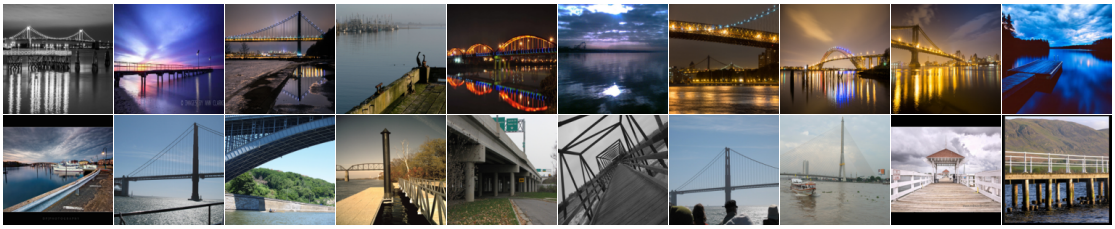
D.4. Comparing Versions of Stable Diffusion

Text-to-image generation details. We use the Huggingface models "CompVis/stable-diffusion-v1-4" and "stabilityai/stable-diffusion-2-1" with guidance of 7.5 and

(a) Wig Diffs: “Close up views of dolls”, “Japanese style dolls”, “Images including Barbie dolls”

(b) Hand-held Computer Diffs: “Apps like Twitter and Whatsapp”, “Digital devices with green screen”, “Interconnection between laptop and smart phone”

(c) Palace Diffs: “East Asian architecture”, “Images featuring the Forbidden City in Beijing”, “Images including red buildings with Chinese writing”



(d) Pier Diffs: “Body of water at night”,“Urban night skyline”, “Long exposure shots”

(e) Schnauzer Diffs: “Black dogs in different settings”, “Terrier puppies with objects”, “Interaction with different objects”

(f) Horizontal Bar Diffs: “Men’s gymnastics events”, “London 2012 Olympics”, “Gymnastics event in 2013”

Figure 12. ImageNetV2 vs. ImageNet. All V2 images are shown in the first row while V1 images are shown in the second row. We show the class name and top 3 difference descriptions generated by VisDiff.

(a) Tobacco Shop Diffs: “Sign hanging from the side of a building”, “Falcon images”,“Presence of street signs”

(b) Digital Watch Diffs: “Watches displayed in a group”,“Arrangement of multiple watches”,“Watches with colored straps”

(c) Missile Diffs: “People posing with large missiles”,“people posing with rockets”,“missiles on display”

(d) Pot Pie Diffs: “Comparison of food size to coins”,“Utilization of cutting board”,“Inclusion of beverages”

(e) Toy Shop Diffs: “People shopping in store”,“Female customer in store”,“Specific location based scene”

Figure 13. CLIP vs. ResNet. All CLIP correctly classified but ResNet incorrectly classified images are shown in the first row while other images are shown in the second row. We show the class name and top 3 difference descriptions generated by VisDiff.

negative prompts “bad anatomy, bad proportions, blurry, cloned face, cropped, deformed, dehydrated, disfigured, duplicate, error, extra arms, extra fingers, extra legs, extra limbs, fused fingers, gross proportions, jpeg artifacts, long neck, low quality, lowres, malformed limbs, missing arms, missing legs, morbid, mutated hands, mutation, mutilated, out of frame, poorly drawn face, poorly drawn hands, signature, text, too many fingers, ugly, username, watermark, worst quality”.

VisDiff details. Unlike the previous applications, there exists a one-to-one mapping between $\mathcal{D}_A$ and $\mathcal{D}_B$ through the generation prompt. Therefore, we modify the subset sampling process to include the images generated from the same prompts and modify the proposer’s prompt to include the generation prompts (Figure 14). We used LLaVA-1.5

Diffusion Text-based Proposer Prompt

The following are the result of captioning two groups of images generated by two different image generation models, with each pair of captions corresponding to the same generation prompt:

Prompt: red apples on a tree with green leaves
Group A: a tree filled with red apples hanging from its branches. There are a total of nine apples visible in the scene, with some of them appearing to be ripe and ready to be picked. The apples are arranged in various positions on the tree, with some closer to the top and others near the bottom. The tree appears to be a healthy and thriving source of fresh fruit.
Group B: a tree filled with a variety of apples hanging from its branches. There are several apples of different sizes and colors, including red and green, scattered throughout the tree. Some apples are hanging higher up, while others are closer to the lower branches. The tree appears to be a lush, healthy source of fresh fruit.
....

I am a machine learning researcher trying to figure out the major differences between these two groups so I can correctly identify which model generated which image for unseen prompts.

Come up with 10 distinct concepts that are more likely to be true for Group A compared to Group B. Please write a list of captions (separated by bullet points "*") . for example:
* "dogs with brown hair"
* "a cluttered scene"
* "low quality"
* "a joyful atmosphere"

Do not talk about the caption, e.g., "caption with one word" and do not list more than one concept. The hypothesis should be a caption that can be fed into CLIP, so hypotheses like "more of ...", "presence of ...", "images with ..." are incorrect. Also do not enumerate possibilities within parentheses. Here are examples of bad outputs and their corrections:
* INCORRECT: "various nature environments like lakes, forests, and mountains" CORRECTED: "nature"
* INCORRECT: "images of household object (e.g. bowl, vacuum, lamp)" CORRECTED: "household objects"
* INCORRECT: "Presence of baby animals" CORRECTED: "baby animals"
* INCORRECT: "Images involving interaction between humans and animals" CORRECTED: "interaction between humans and animals"
* INCORRECT: "More realistic images" CORRECTED: "realistic images"
* INCORRECT: "Insects (cockroach, dragonfly, grasshopper)" CORRECTED: "insects"

Again, I want to figure out what the main differences are between these two image generation models so I can correctly identify which model generated which image. List properties that hold more often for the images (not captions) in group A compared to group B. Answer with a list (separated by bullet points "*"). Your response:

Figure 14. Modified proposer's prompt for StableDiffusion analysis.

for captioning rather than BLIP-2 because we were particularly interested in the details of the images.

Top differences. Top 5 differences are shown in Table 10.

More True for SDv2	AUROC Parti	AUROC DiffDB
colorful and dynamic collages of shapes or items	0.70	0.71
vibrant colors	0.72	0.70
strong contrast in colors	0.68	0.68
reflective surfaces	0.68	0.68
artworks placed on stands or in frames	0.64	0.66

Table 10. Concepts more true for SDv2 than v1. Differences are proposed by running VisDiff on PartiPrompts images. These differences obtain similar scores on the unseen DiffusionDB images, indicating that these differences generalize to various prompts.

Visualizations. We provide 50 random samples of SDv2 and SDv1 images generated with DiffusionDB prompts in Figure 15. These examples clearly verify that SDv2-generated images contain more vibrant contrasting colors and artwork or objects in frames or stands.

Edge analysis. One interesting finding from VisDiff is that SDv2 generated images contain more image frames than SDv1, such as a white border characterized by thick, straight lines spanning much of the image. To quantify this, we employed a Canny edge detector and searched for straight white lines in the images, with a thickness ranging from 5 to 20 pixels and a length exceeding 300 pixels (given the image size is 512x512). Applying this analysis to DiffusionDB images revealed that 13.6% of SDv2 images exhibited such lines, as opposed to only 5.4% from SDv1. This statistic provides additional evidence for such difference.

D.5. Memorable Images

Top differences. The top 25 difference descriptions generated by VisDiff are presented in Table 11.

Classification analysis. To validate whether the generated differences for memorable and forgettable images make sense, we use CLIP to classify each image in the LaMem dataset to these 25+25 differences and then assign the label "forgettable" or "memorable" based on where the difference is from. For example, if an image has the highest cosine similarity with "close-up of individual people",

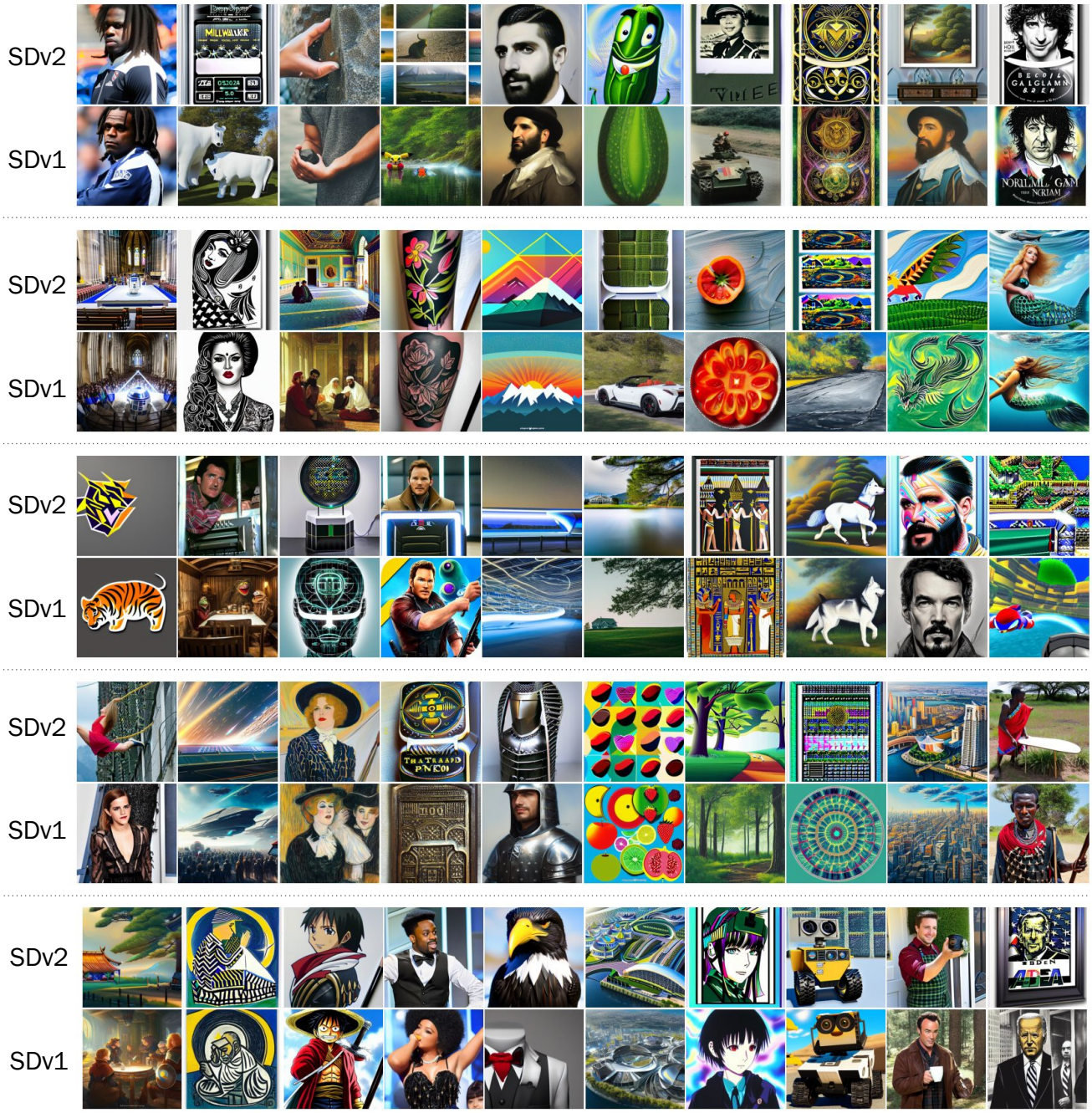


Figure 15. Randomly sampled images generated from SDv2 and v1 using DiffusionDB prompts.

we assign its label as "memorable". We observe a 89.8% accuracy on the LaMem test set, demonstrating that these differences provide strong evidence to classify whether images are memorable or forgettable.

E. Failure Cases and Limitations

In this section, we summarize the failure cases and limitations of VisDiff algorithm.

E.1. Caption-based Proposer

While our evaluation in the main paper shows that the caption-based proposer outperforms other counterparts by

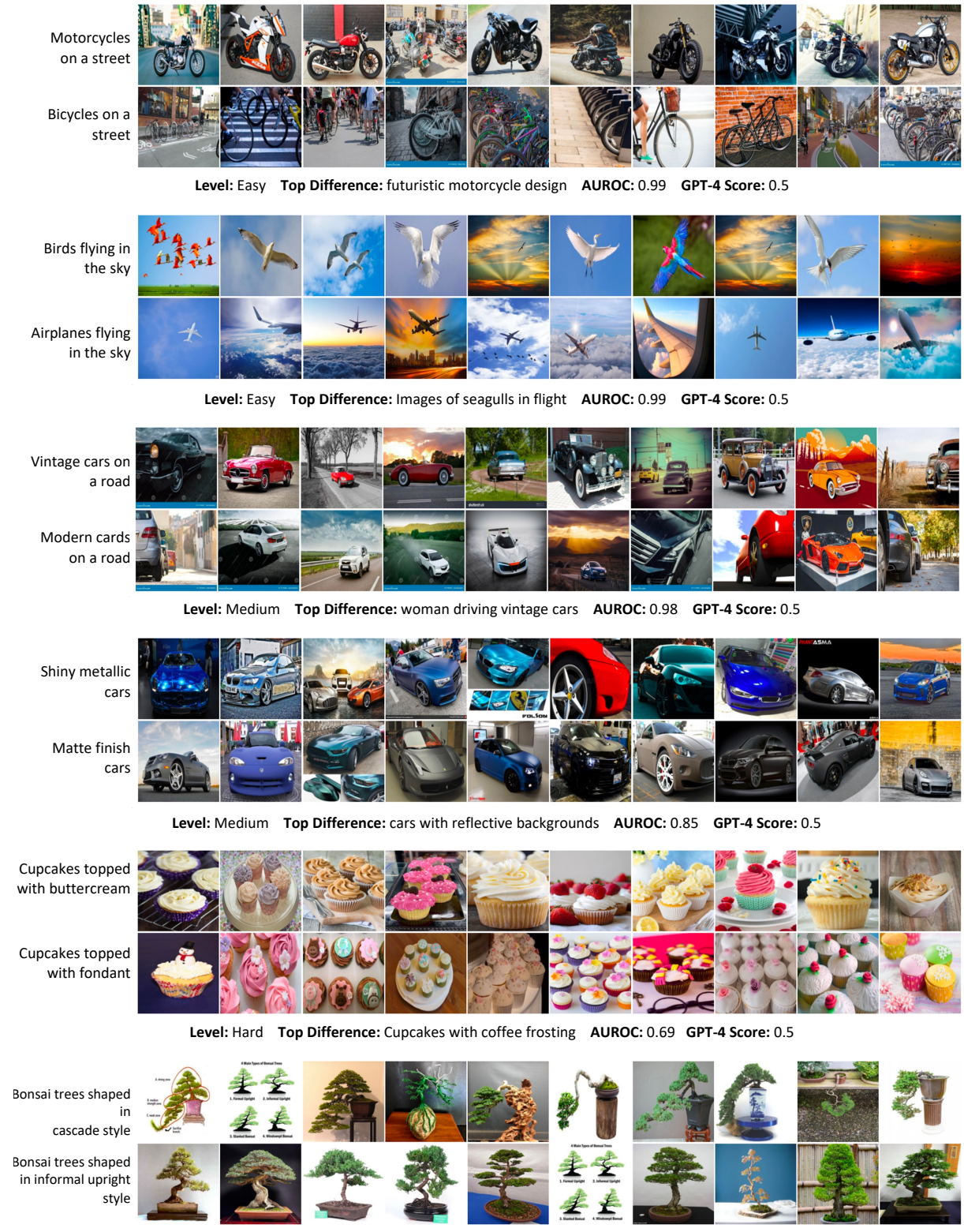


Figure 16. PairedImageSets examples where VisDiff fails. We show the ground-truth difference, top difference predicted by VisDiff, AUROC score output by the ranker, and evaluation of the predicted difference by GPT-4.

Memorable	close-up of individual people, use of accessories or personal items, tattoos on human skin, close-up on individuals, humorous or funny elements, artistic or unnaturally altered human features, humorous elements, detailed description of tattoos, fashion and personal grooming activities, pop culture references, collectibles or hobbies, light-hearted or humorous elements, themed costumes or quirky outfits, animated or cartoonish characters, emphasis on fashion or personal style, close-up of objects or body parts, close-up facial expressions, unconventional use of everyday items, images with a playful or humorous element, focus on specific body parts, silly or humorous elements, people in casual or humorous situations, detailed description of attire, quirky and amusing objects, humorous or playful expressions
Forgettable	Sunsets and sunrises, serene beach settings, sunset or nighttime scenes, agricultural fields, clear daytime outdoor settings, landscapes with water bodies, images captured during different times of day and night, Beautiful skies or sunsets, abandoned or isolated structures, natural elements like trees and water, urban cityscapes, urban cityscapes at night, various weather conditions, Afar shots of buildings or architectural structures, outdoor landscapes, cityscapes, Cityscapes and urban environments, Scenic outdoor landscapes, landscapes with mountains, Picturesque mountain views, expansive outdoor landscapes, Scenic landscapes or nature settings, Serene and tranquil environments, scenic landscapes, scenes with a serene and peaceful atmosphere

Table 11. Top 25 differences for memorable and forgettable images.

a large margin, translating images to captions may lead to information loss. For example, as shown in Figure 16, fine-grained differences between groups "Cupcakes topped with buttercream" and "Cupcakes topped with fondant" is overlooked due to generic captions. We expect using captioning prompts tailored to the application domain can mitigate this issue.

Furthermore, despite providing task context and several in-context examples, we noted instances where GPT-4 predominantly focused on the captions rather than the underlying high-level visual concepts. A frequent error involves generating concepts related more to the caption than the image, such as "repetition of 'bonsai' in the caption," as illustrated in Figure 16. We anticipate that this issue will diminish as LLMs' instruction-following ability improves.

E.2. Feature-based Ranker

Several of VisDiff's ranker failure cases stem from biases and limitations in CLIP. For example, nuanced differences such as "a plant to the left of the couch" are often assigned lower rankings because CLIP struggles with precise location details, and minor variations in phrasing can lead to significant differences in similarity scores.

Additionally, using AUROC on cosine similarities as a ranking metric is sensitive to outliers in cosine similarity scores. In practice, we have noticed that outliers can cause very specific difference descriptions to be scored higher than more general differences. For instance, as shown in Figure 16, with $\mathcal{D}_A$ being "Birds flying in the sky" and $\mathcal{D}_B$ "Airplanes flying in the sky," the hypothesis "Images of seagulls in flight" received a higher AUROC score than the more broadly applicable "birds in flight".

E.3. LLM-based Evaluation

As demonstrated in the main paper, large language models generally align well with human evaluations. However, there are instances where they fail to accurately score differences against the ground truth descriptions. An example from VisDiffBench involves the description "Green apples in a basket" for $\mathcal{D}_A$ and "Red apples in a basket" for $\mathcal{D}_B$. Here, the top hypothesis by VisDiff, "Green apples" received a score of only 0.5 instead of the expected 1.0. These errors are expected to diminish as LLM improves.

E.4. VisDiffBench

Most differences in VisDiffBench focus on objects, styles, and actions. Differences such as object position, size, or image quality are missing. Additionally, since PairedImageSets is compiled by scraping images from the web, the datasets inevitably include noise. For instance, searching for "a cat to the left of a dog" often yields images with a cat on the right instead.

E.5. Reliance on Large Pre-trained Models

Our approach is fundamentally based on large, pre-trained vision-language foundation models. These models' extensive capabilities make them adaptable for a variety of tasks. However, inherent biases and limitations in these models may be transferred to our method. Additionally, these models might be confined to domains observed during pre-training, potentially limiting their applicability to novel domains, such as biomedical imaging. Nevertheless, we anticipate that rapid advancements in foundation model development will mitigate these issues, thereby enhancing our method's effectiveness.