Appendices

All codes, data, and instructions for our CompBench can be found in https://github.com/RaptorMai/CompBenchReview. CompBench is released under a Creative Commons Attribution 4.0 License (CC BY 4.0).

Our supplementary materials are summarized as follows:

- Appendix A: Limitations, social impacts, ethical considerations, and license of assets.
- Appendix B: CompBench curation details (cf. §4.2 and §5.1 in the main text).
- Appendix C: Training details on LLaVA-1.6 (cf. §5.3 in the main text).
- Appendix D: More qualitative examples.

A Discussions

A.1 Limitations

While we conducted a human evaluation study to establish the upper bound performance on CompBench, the study is currently limited to 140 samples assessed by five evaluators (cf. §5.3 in the main text). We plan to expand the study to a larger scale in future work.

A.2 Social impacts

CompBench evaluates the comparative reasoning abilities of MLLMs in images. A potential negative impact of our work is that malicious users might exploit our concept (i.e., comparison) to compare ethical or offensive content. Therefore, it is essential to incorporate effective safeguards in MLLMs to filter out any inappropriate materials.

A.3 Ethical considerations

All fourteen datasets (cf. Table 1 in the main text) that we used to curate CompBench adhere to strict guidelines to exclude any harmful, unethical, or offensive content. Additionally, we instruct human annotators to avoid generating any personally identifiable information or offensive content during our annotation process. Finally, we do not conduct any study to compare harmful, ethical, or offensive content between the two images.

A.4 License of assets

All fourteen datasets are publicly available, and Table 1 details the licensing information for the assets in each dataset. We release our CompBench under a Creative Commons Attribution 4.0 License (CC BY 4.0) to enhance global accessibility and foster innovation and collaboration in research.

B CompBench Curation Details

B.1 Annotation Details

We create UI interfaces for annotation using Python in Jupyter Notebook and store the annotations in JSON files. In the following sections, we provide detailed descriptions of the annotation process for each dataset, which are omitted in the main text.

MagicBrush [18] is a large-scale, manually annotated dataset for instruction-guided real image editing. For each image, MagicBrush utilizes DALL-E 2 [13] to generate an edited version of the image based on language instructions, such as "let the flowers in the vase be blue." Our goal is to identify pairs of similar images. We thus use CLIP [12] to evaluate the visual similarity between the

Public Dataset	License
MIT-States [5]	N/A
Fashionpedia [7]	CC BY 4.0
VAW [11]	Adobe Research License
CUB-200-2011 [16]	CC BY
Wildfish++ [20]	N/A
MagicBrush [18]	CC BY 4.0
Spot-the-diff [6]	N/A
CelebA [10]	Research-only, non-commercial
FER-2013 [3]	N/A
SoccerNet [2]	MIT License
CompCars [17]	Research-only, non-commercial
NYU-Depth V2 [14]	N/A
VQAv2 [4]	CC BY 4.0
Q-Bench2 [19]	N/A

Table 1: **License of Assets**.

original and edited images. Only pairs exceeding a predetermined similarity threshold are selected as candidate samples for our COMPBENCH. For each selected pair, we then construct a multiple-choice question to ask the difference between two images in the pairs. Concretely, we first use GPT-4V [1] to extract all relevant objects and their attributes from the edited image with the following prompt:

> "Please extract as many components as possible from the provided images. The following examples illustrate some potential components, but the list is not exhaustive. Only provide the component names, separated by commas. If a human or an animal is shown in the images and features such as hair, eyes, hands, mouth, ears, and legs are visible, ensure to include them. Similarly, try to identify all components in as much detail as possible.
>
> Examples of components: leg, eye, ear, food, pillow, flower, plate, window, door, chair, dining table, sofa, banana, bowl, sugar, blender, berry, lizard, watermelon, motorcycle, apple, curtain, cookies, cake, hair, hat, dresses, bacon, butter, jam, bread, surfboard, t-shirt, pants, hands, fridge, plants, cabinet, sink, car, girl, boy."

We treat objects and their attributes (if found) as options for the questions. However, GPT-4V [1] may not capture all relevant objects (options) in the images. We thus request human annotators to add as many relevant options as possible. Finally, annotators are required to select the obvious difference between two images as the correct answer among options and verify the quality of the generated samples (Figure 1).

Spot-the-diff [6] offers video-surveillance image pairs from outdoor scenes, along with descriptions and pixel-level masks of their differences. Similar to MagicBrush, we aim to construct a multiple-choice question to find the obvious difference between the two images. We first prompt the text-only GPT-4 to extract the potentially correct objects from the descriptions of the differences using the following prompt:

> "These sentences describe the differences between the two images. Extract the objects from these sentences. for example, ["there are more people", "the car moved"], you should return "people, car". Please only provide the answer without any explanation and separate the answer names by commas."

Given the extracted objects and the images, GPT-4V is tasked with finding relevant options in the images based on the following prompt:

> "Please list all the objects and attributes associated with the image, for example, black cars, people, trees, white trucks, and yellow poles. Only provide one attribute (adjective) per object. Please only provide the answer without any explanation

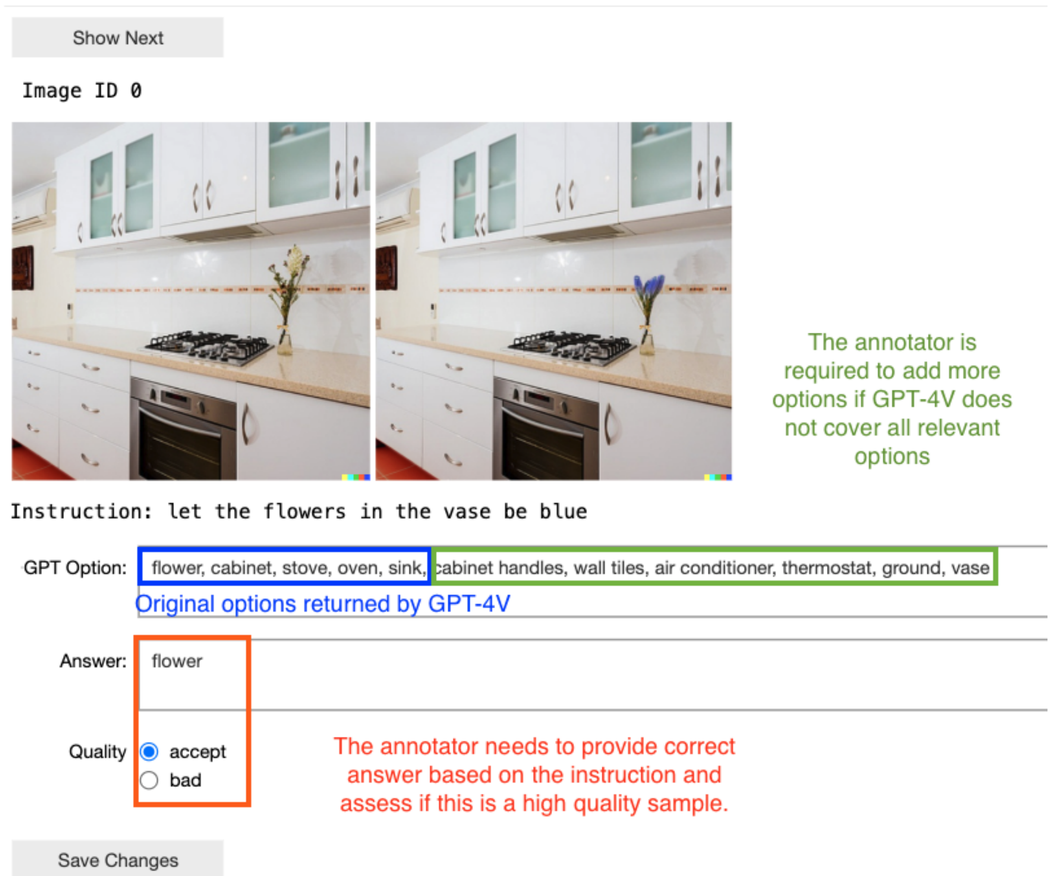

Figure 1: **Annotation Interface for MagicBrush.**

> and separate the answer names with commas. Ensure to include these objects: [OBJECTS FROM LAST STEP]"

We then instruct human annotators to include additional options (if necessary) and identify the most evident difference between two images from the available options as the correct answer (Figure 2).

MIT-States [5] includes 245 objects with 115 visual attributes or states from online sources such as food or device websites. Each folder in this dataset is named by (adjective, noun), e.g., tall tree, where the adjective describes the state or the attributes and the noun is the object. All the images in this folder share the same adjective and noun. We apply rule-based approaches to generate questions about relative degrees of attributes or states between objects (e.g., "Which tree is taller?"). We then present the questions with the corresponding images in this folder to annotators. The annotators are tasked to select pairs from all the images, label the correct answers (binary: left/right), and filter out any irrelevant or nonsensical questions about the images. In addition, the annotators are required to determine the attribute or state types by selecting from the following options: Size, Color, Texture, Shape, Pattern, State, or None. We filter out examples where the type or answer is None. The annotation UI interface is shown in Figure 3.

VAW [11] provides a large-scale collection of 620 unique attributes, including color, shape, and texture. We process VAW in the same manner as MIT-States, as detailed in Figure 3.

CUB-200-2011 [16] catalogs 15 bird parts and their attributes (e.g., "notched tail"). We group images by species with the same attributes (e.g., "curved bill") and extract visually similar image pairs from each group. We then prompt GPT-4 to transform visual attributes into questions that compare them using the following in-context prompt:

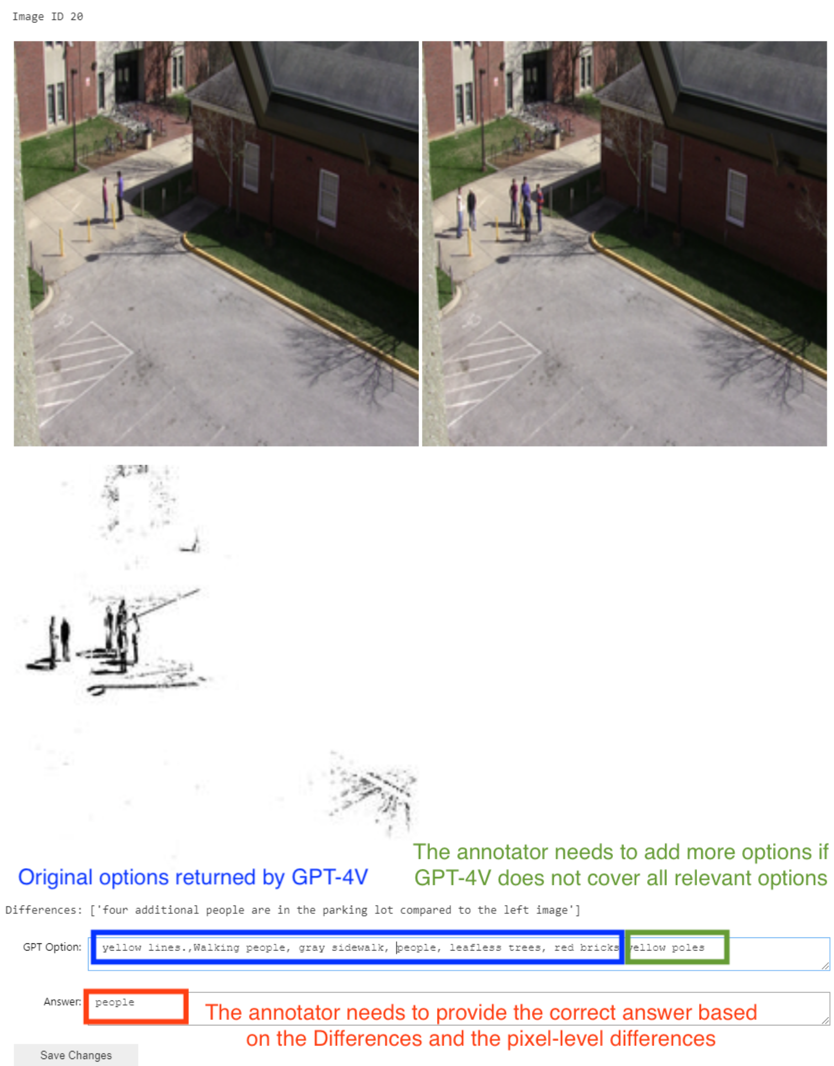

Figure 2: **Annotation Interface for Spot-the-diff.**

> "I want to turn some text describing the attributes of birds into a question comparing these attributes between birds in two different images. Here are some examples: Attribute: has_bill_shape::hooked, Questions: Which bird has a more hooked bill? Attribute: has_crown_color::brown, Questions: Which bird has more brown on its crown?
>
> Please turn this list of attributes into these questions in this format or style. I want a dictionary format output. [ATTRIBUTE LIST]"

The annotators receive all images in each group along with corresponding comparative questions generated by GPT-4. They are asked to select the pairs from the images and label the correct answers (binary: left/right). The annotation interface is shown in Figure 4.

Wildfish++ [20] details 22 characteristics (e.g., "brown pelvic fins") of various fish species and provides detailed descriptions of the differences between two visually similar species. Using the characteristics and the descriptions of difference, we first ask annotators to generate comparative questions (e.g., "Which fish has lighter brown pelvic fins?"). Subsequently, we pass all images from the two similar species along with the corresponding question to the annotators. They select one image from each group to form a pair and label the correct answers as either left or right (Figure 5).

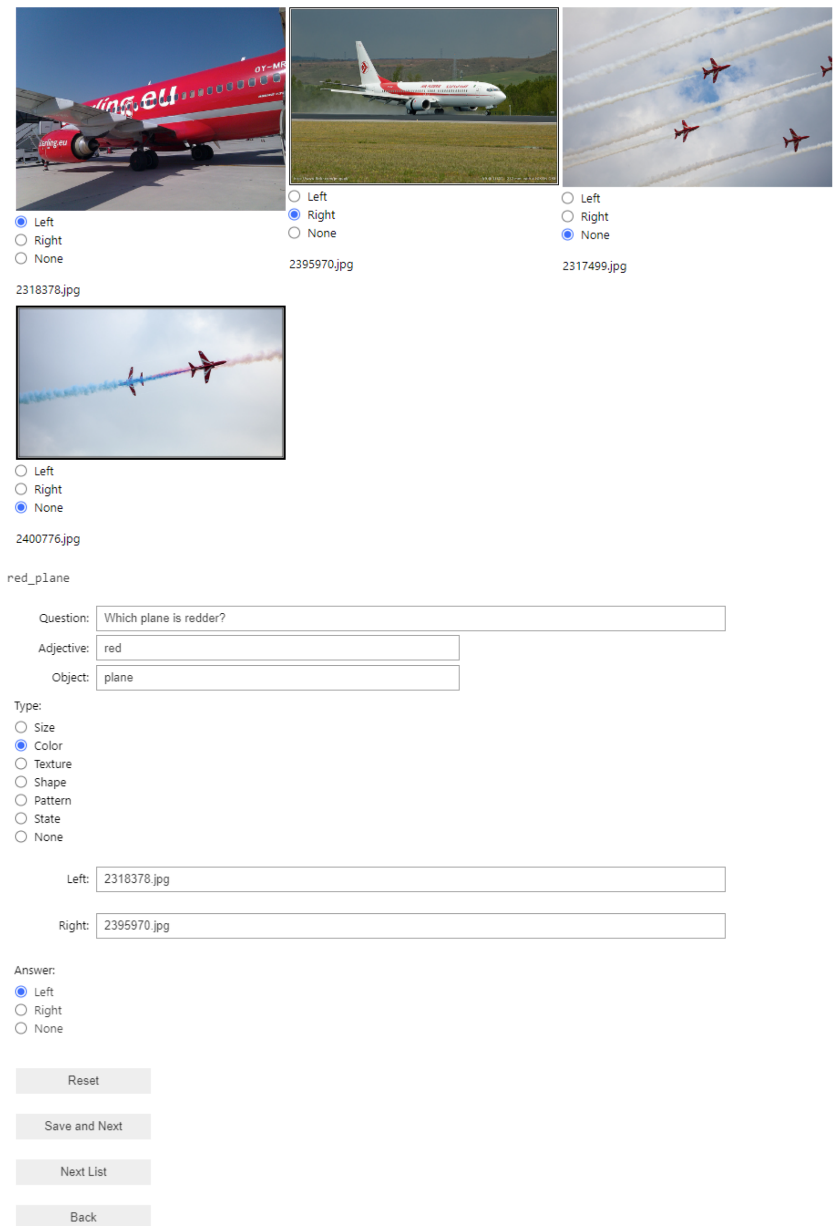

Figure 3: **Annotation Interface for MIT-States and VAW.**

Fashionpedia [7] is tailored to clothing and accessories and contains 27 types of apparel along with 294 detailed attributes. We group images by (attribute, type), e.g., square neckline. We apply rule-based approaches to generate questions about relative degrees of attributes (e.g., "Which neckline is more square?") for each group. We then present images of the same type with different attributes, such as "square neckline" and "oval neckline" to the annotators. The annotators are required to select one image from each group to form a pair, choose one between questions from two attributes, and label the correct answer (binary: left/right). The annotation UI interface is shown in Figure 6.

NYU-Depth V2 [14] features indoor scenes with object segments and depths. Using the segmentation maps, we identify objects within each image and group images containing the same objects. We apply rule-based approaches to generate questions about spatial relative comparisons (e.g., "Which [OBJECT] is closer to the camera?"). The annotator needs to select pairs from all the images in the same group and label the correct answers either left or right (Figure 7).

CelebA [10] is a large-scale facial attributes dataset featuring over 200K celebrity images, each annotated with 40 attributes. We focus on images labeled with the "smiling" attribute, as it is the only attribute related to the emotion in the dataset. We generate a comparative question such as "Which person smiles more?". The annotators are tasked with selecting pairs from all images with the smiling attribute and labeling the correct answers either left or right (Figure 8).

FER-2013 [3] contains grayscale images along with categories describing the emotion of the person, including Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. We leverage rule-based approaches to generate questions about relative emotional comparisons (e.g., "Which person looks more [EMOTIONAL ADJECTIVE]?"). The annotators are required to select pairs from images that share the same emotional attribute and determine the correct answers as either left or right (Figure 9).

SoccerNet [2], CompCars [17], VQAv2 [4], Q-bench2 [19] are automatically processed to generate samples for CompBench using their metadata and CLIP visual similarity. For more details, please refer to §4.2 of the main text.

B.2 Language Prompts for MLLMs

Table 2 summarizes our language prompts for evaluating MLLMs. We observe that in the case of SoccerNet [2], Gemini1.0-pro [15] always predicts the answer "Left" for binary questions (e.g., "These are two frames related to [SOCCER_ACTION] in a soccer match. Which frame happens first? Please only return one option from (Left, Right) without any other words."). We thus prompted the Gemini to answer open-ended questions (as shown in Table 2) instead. We then task human evaluators with verifying whether its responses (i.e., textual descriptions) match the ground-truth answers to calculate its performance. For a fair comparison, we apply the same open-ended questions to other models (i.e., GPT-4V [1], LLaVA-1.6 [9], VILA-1.5 [8]) and report their accuracies.

B.3 Model Evaluation

We use official APIs to evaluate proprietary MLLMs, GPT-4V [1] and Gemini [15]. For GPT-4V, we use the version of gpt-4-turbo[1]. For Gemini, we use the Gemini1.0 Pro Vision[2]. For open source models such as LLaVa-1.6-34b [9][3] and VILA-1.5-40b [8][4], we utilize their official source codes and conduct inference on NVIDIA RTX 6000 Ada GPUs.

B.4 Human Annotators & Evaluators

We recruited five in-house human annotators from our research team to work on CompBench. The annotators are instructed to avoid generating any personally identifiable information or offensive content during the annotation process. Furthermore, we recruited another five human evaluators, who were not involved in the annotation, to measure the upper bound performance on CompBench. The workloads for annotation and evaluation were distributed equally among annotators and evaluators.

C Training details on LLaVA-1.6

As discussed in §5.3 of the main text, we conduct a study to evaluate whether fine-tuning enhances the comparative capabilities of MLLMs. Concretely, we focus on two relativities: Temporality and Quantity. For temporality, we construct a total of 20.6K training examples from SoccerNet [2], following the similar data collection and annotation protocol described in §4.2.5 of the main text. For quantity, we curate a total training set of 20.9K samples from VQAv2 [4], based on the similar data collection and annotation pipeline in §4.2.7 of the main text. We fine-tune LLaVA-1.6-34b [9] on each of these training datasets separately, using LoRA techniques. We follow similar hyperparameter

[1] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
[2] https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.0-pro-vision
[3] https://github.com/haotian-liu/LLaVA
[4] https://github.com/Efficient-Large-Model/VILA

Dataset	Model	Lagnauge Prompt
ST, FA, VA, CU, WF, CE, FE, ND	GPT-4V LLaVA-1.6 VILA-1.5	"[QUESTION] If you choose the first image, return Left, and if you choose the second image, return Right. Please only return either Left or Right without any other words, spaces, or punctuation."
	Gemini1.0-pro	"[QUESTION] If you choose the first image, return First, and if you choose the second image, return Second. Please only return either First or Second without any other words, spaces, or punctuation."
MB, SD	GPT-4V LLaVA-1.6 VILA-1.5 Gemini1.0-pro	"What is the most obvious difference between the two images? Choose from the following options. If there is no obvious difference, choose None. Options: None, [OPTIONS]. Please only return one of the options without any other words. "
SN	GPT-4V LLaVA-1.6 VILA-1.5 Gemini1.0-pro	"These are two frames related to [SOCCER_ACTION] in a soccer match. Which frame happens first?"
CC	GPT-4V LLaVA-1.6 VILA-1.5	"Based on these images, which car is newer in terms of its model year or release year? Note that this question refers solely to the year each car was first introduced or manufactured, not its current condition or usage. If you choose the first image, return Left, and if you choose the second image, return Right. Please only return either Left or Right without any other words, spaces, or punctuation."
	Gemini1.0-pro	Based on these images, which car is newer in terms of its model year or release year? Note that this question refers solely to the year each car was first introduced or manufactured, not its current condition or usage. If you choose the first image, return First, and if you choose the second image, return Second. Please only return either First or Second without any other words, spaces, or punctuation."
VQ	GPT-4V LLaVA-1.6 VILA-1.5 Gemini1.0-pro	"[QUESTION] If the second image has more, return Right. If the first image has more, return Left. If both images have the same number, return Same. Please only return either Left or Right or Same without any other words, spaces, or punctuation."
QB	GPT-4V LLaVA-1.6 VILA-1.5 Gemini1.0-pro	"[QUESTION] Options: [OPTIONS]"

Table 2: **Language prompts for evaluating MLLMs**. ST: MIT-States [5], FA: Fashionpedia [7], VA: VAW [11], CU: CUB-200-2011 [16], WF: Wildfish++ [20], MB: MagicBrush [18], SD: Spot-the-diff [6], CE: CelebA [10], FE: FER-2013 [3], SN: SoccerNet [2], CC: CompCars [17], ND: NYU-Depth V2 [14], VQ: VQAv2 [4], QB: Q-Bench2 [19].

settings as those provided in the official LLaVA source codes. For instance, batch size/the number of epochs/learning rate are 16/3/2e-5, respectively. See the training script in our GitHub repository for the complete configuration. All models are fine-tuned on four NVIDIA RTX 6000 Ada GPUs.

D More qualitative examples

In addition to the main text, we show more qualitative examples from each of fourteen datasets in Figure 10, Figure 11, Figure 12, Figure 13, and Figure 14. We observe that GPT-4V, one of the leading MLLMs, often faces challenges across a range of relative comparison tasks.

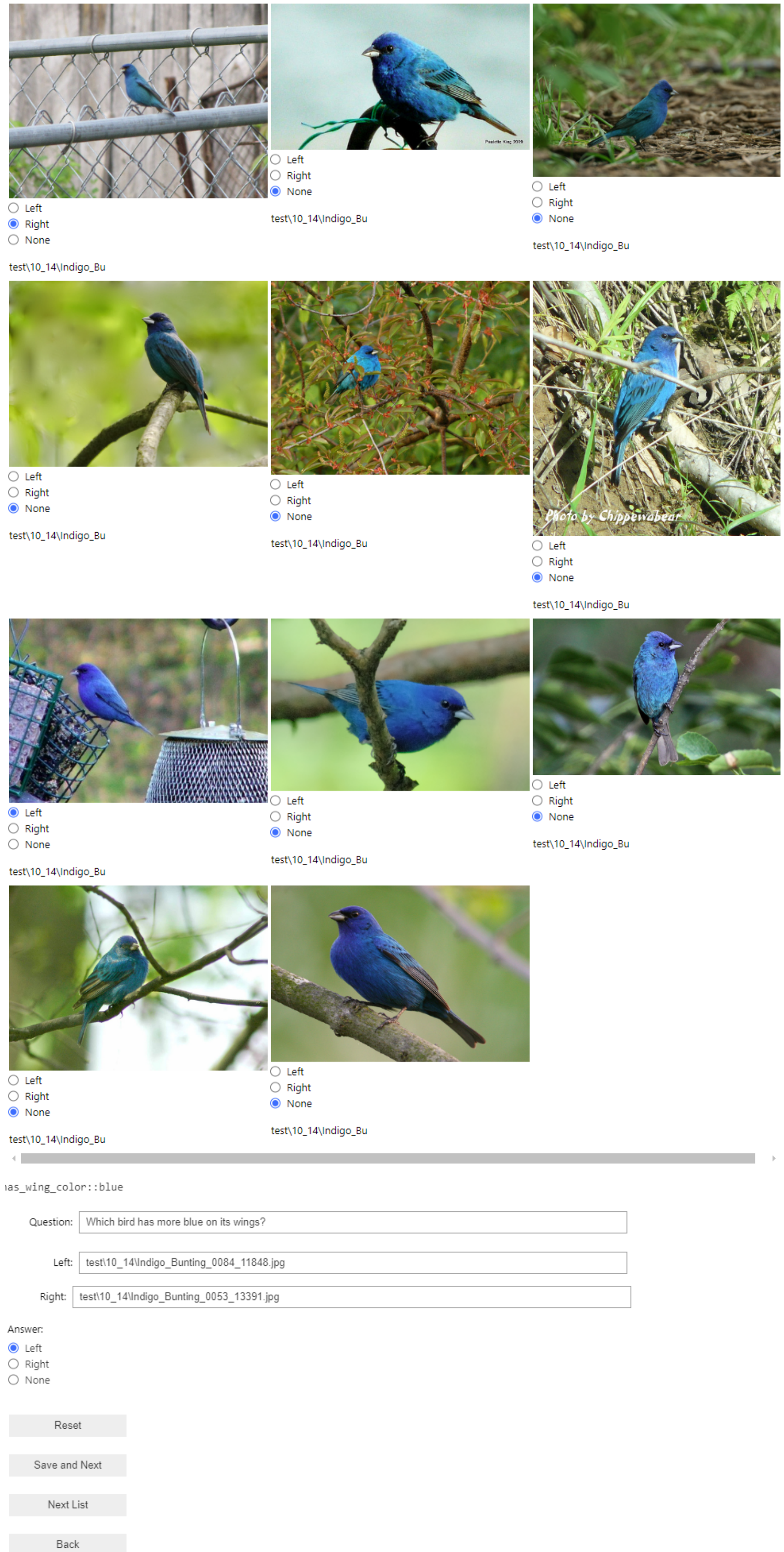

Figure 4: **Annotation Interface for CUB-200-2011.**

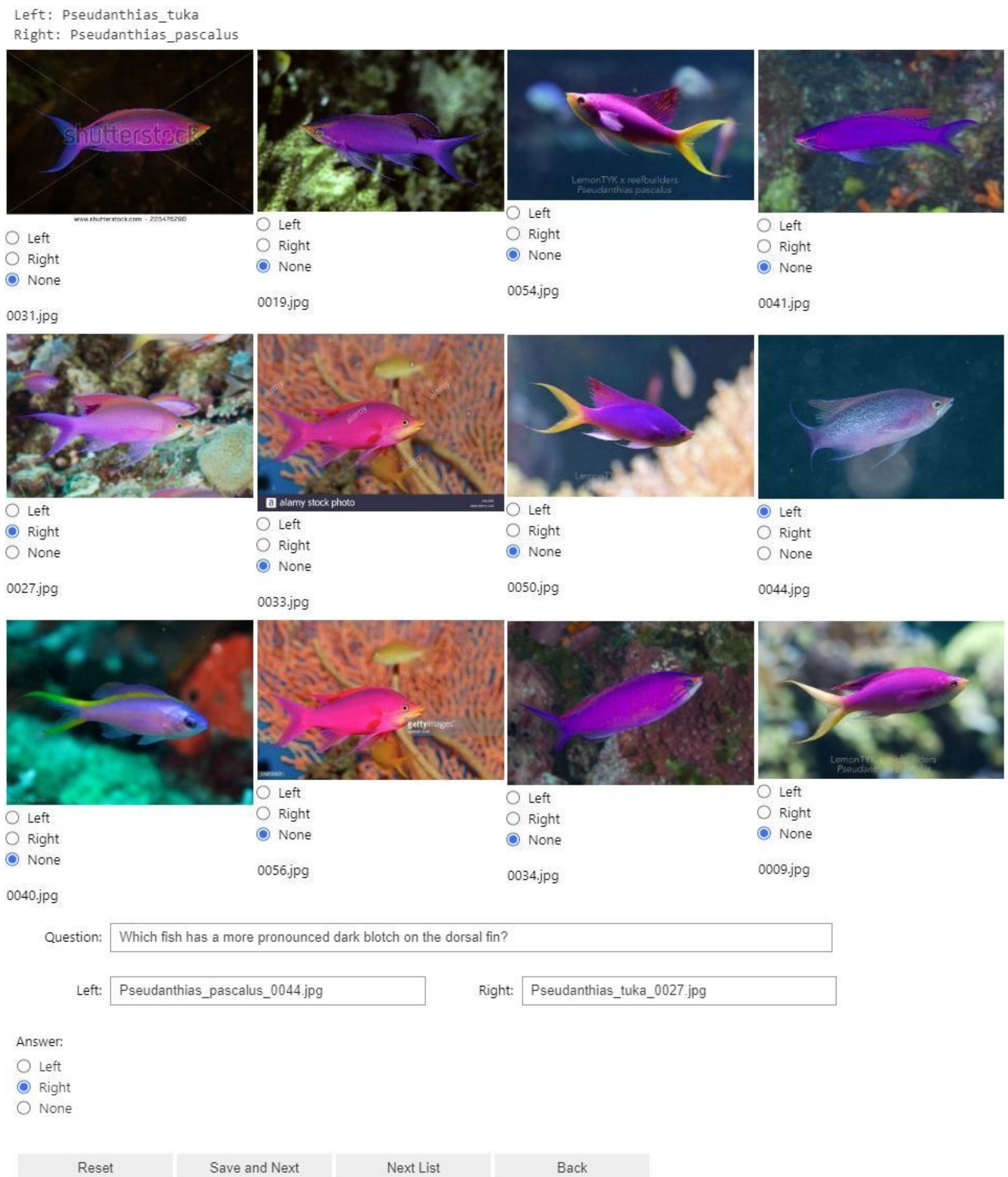

Figure 5: **Annotation Interface for Wildfish++.**

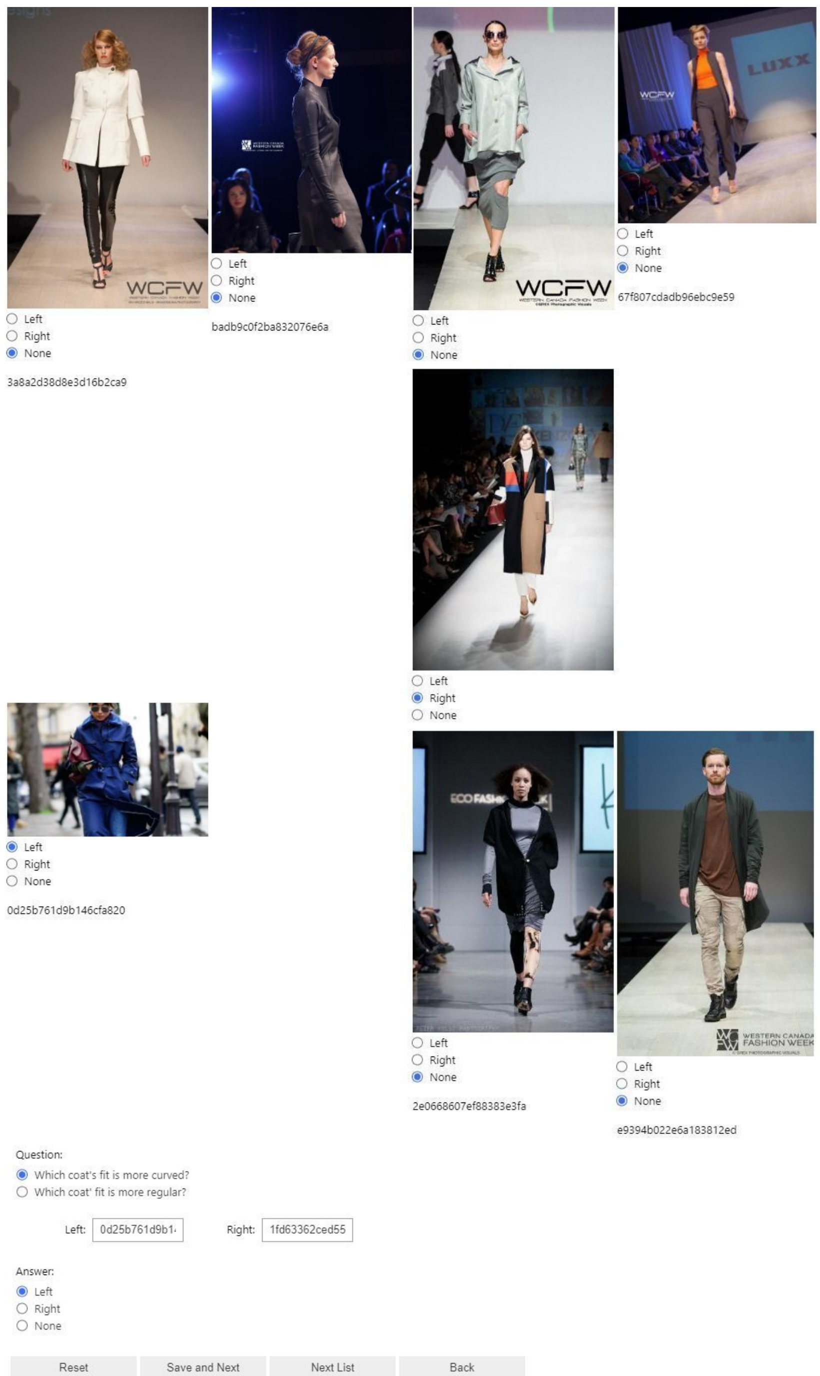

Figure 6: **Annotation Interface for Fashionpedia.**

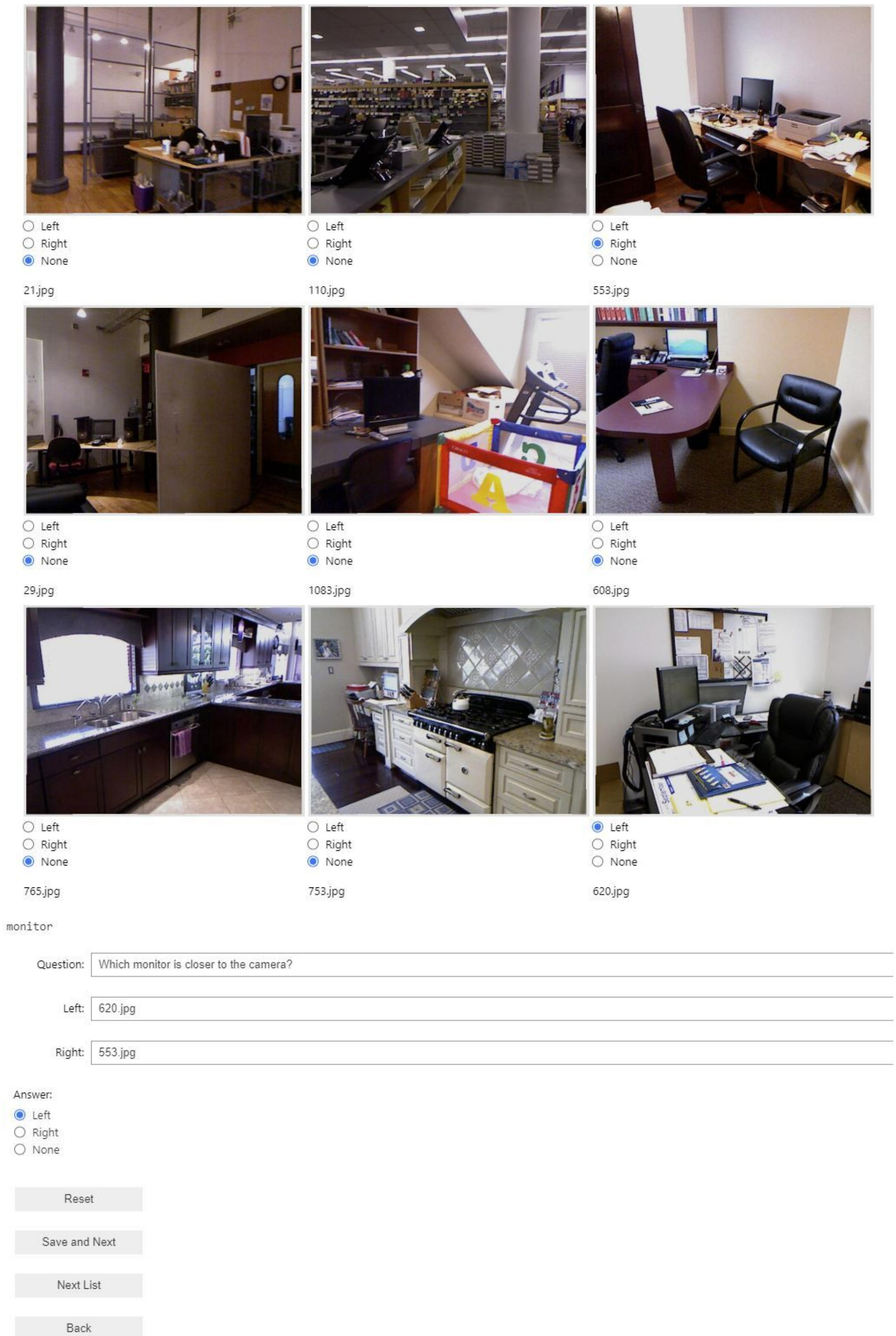

Figure 7: **Annotation Interface for NYU-Depth V2.**

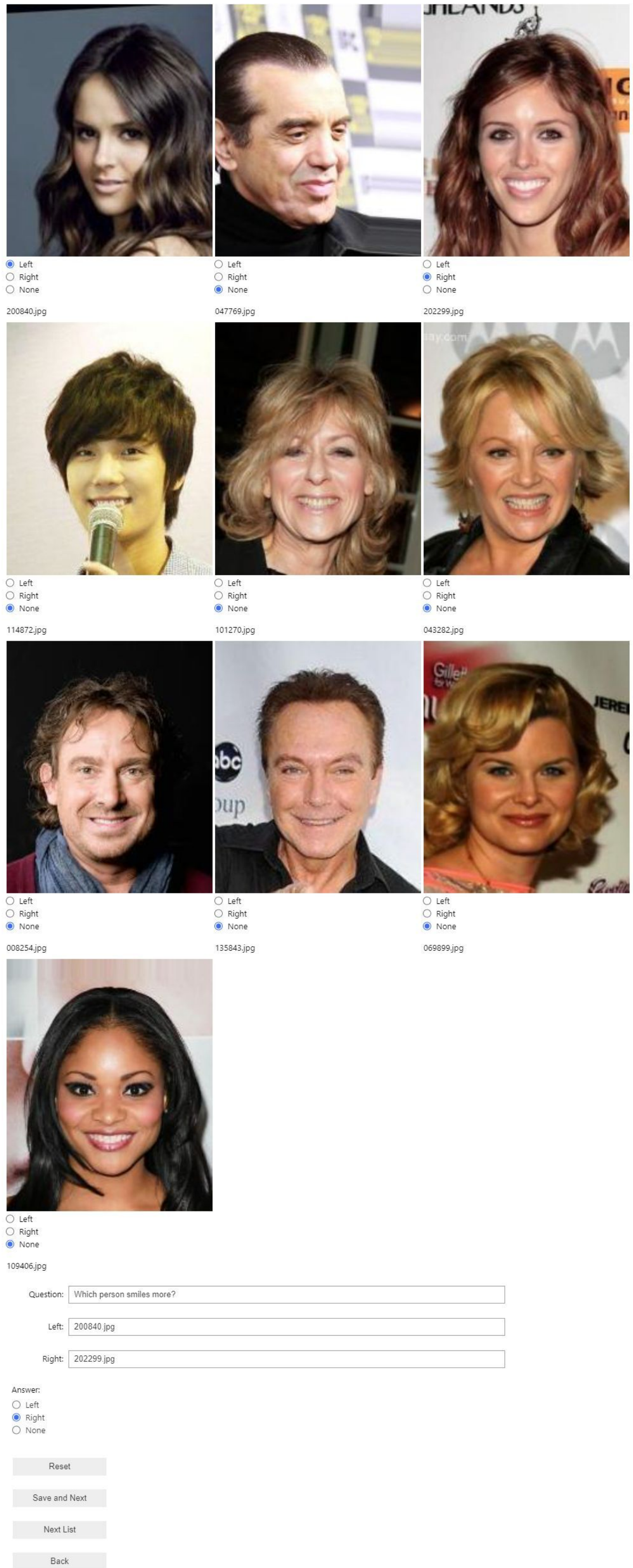

Figure 8: **Annotation Interface for CelebA.**

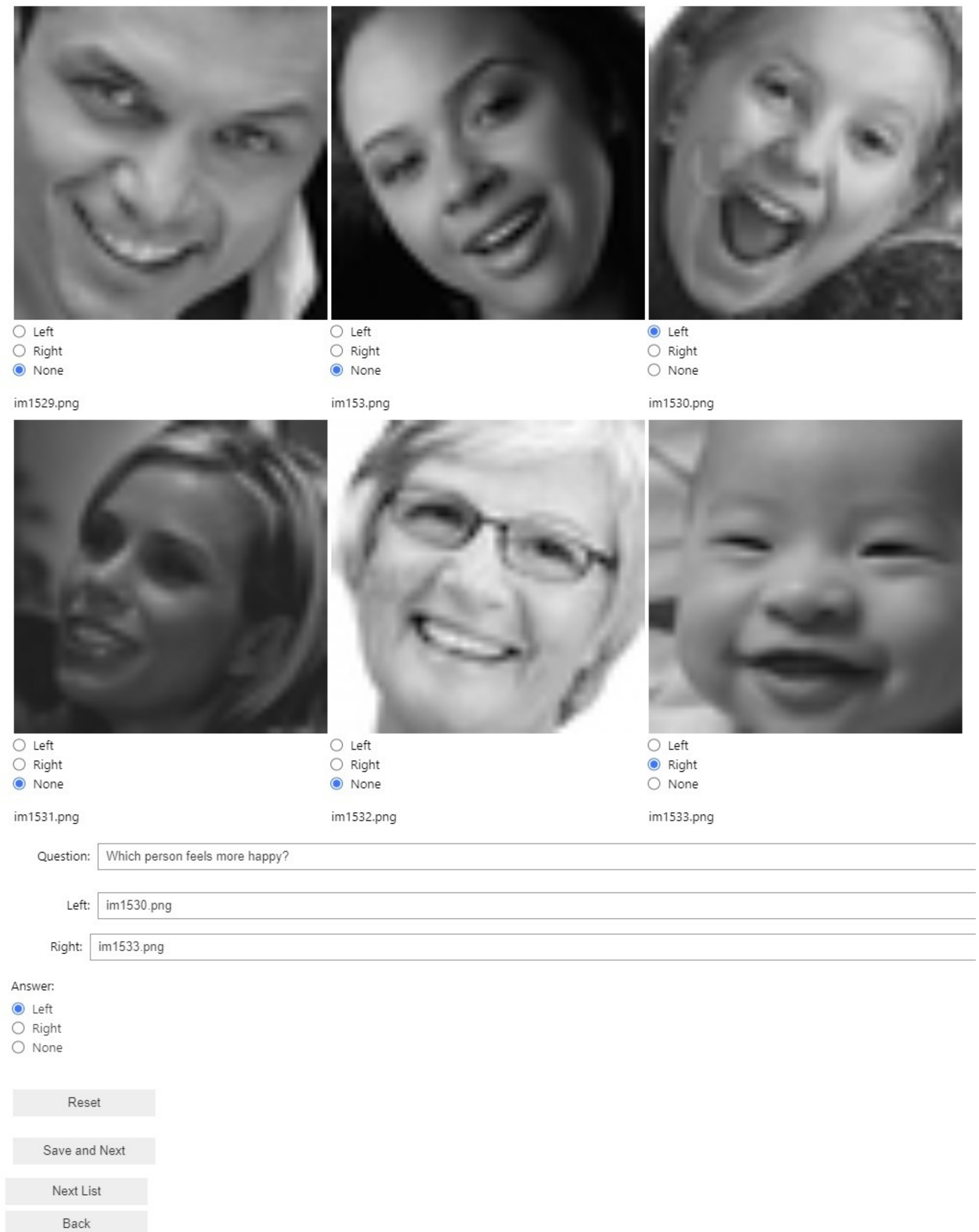

Figure 9: **Annotation Interface for FER-2013.**

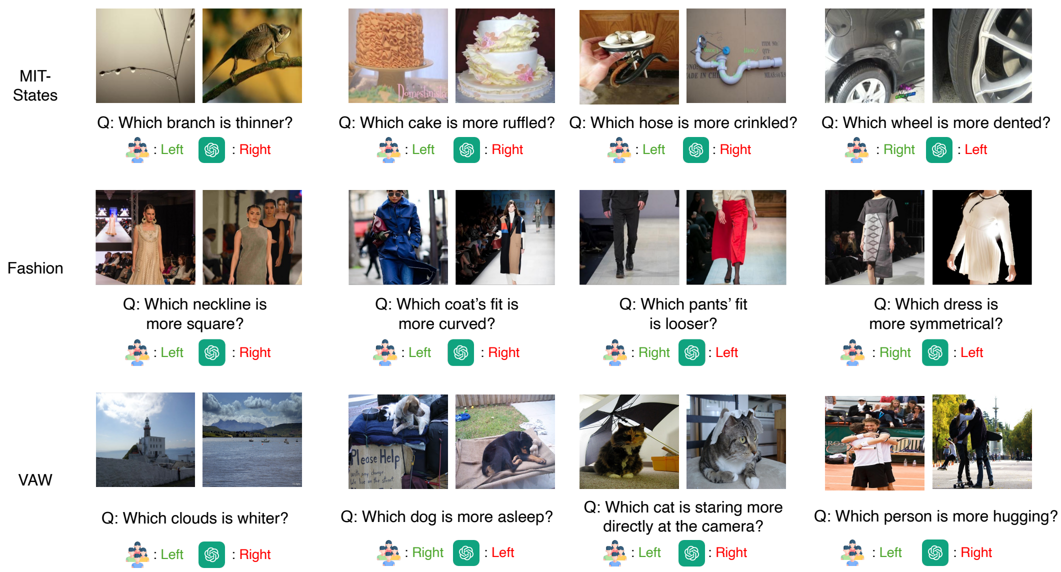

Figure 10: **Qualtiative examples on MIT-States [5], Fashionpedia [7], and VAW [11].**

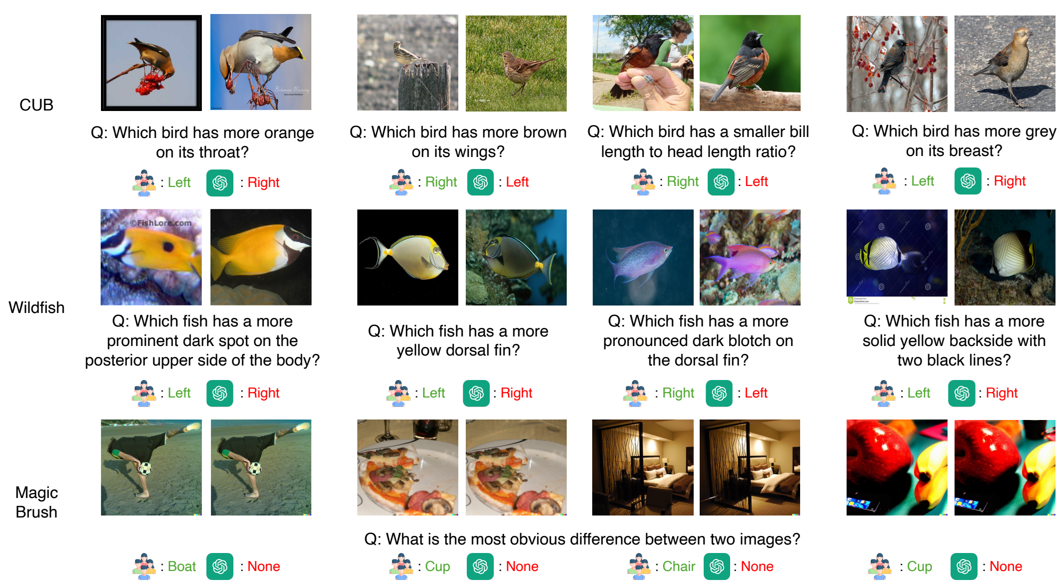

Figure 11: **Qualtiative examples on CUB-200-2011 [16], Wildfish++ [20], and MagicBrush [18].**

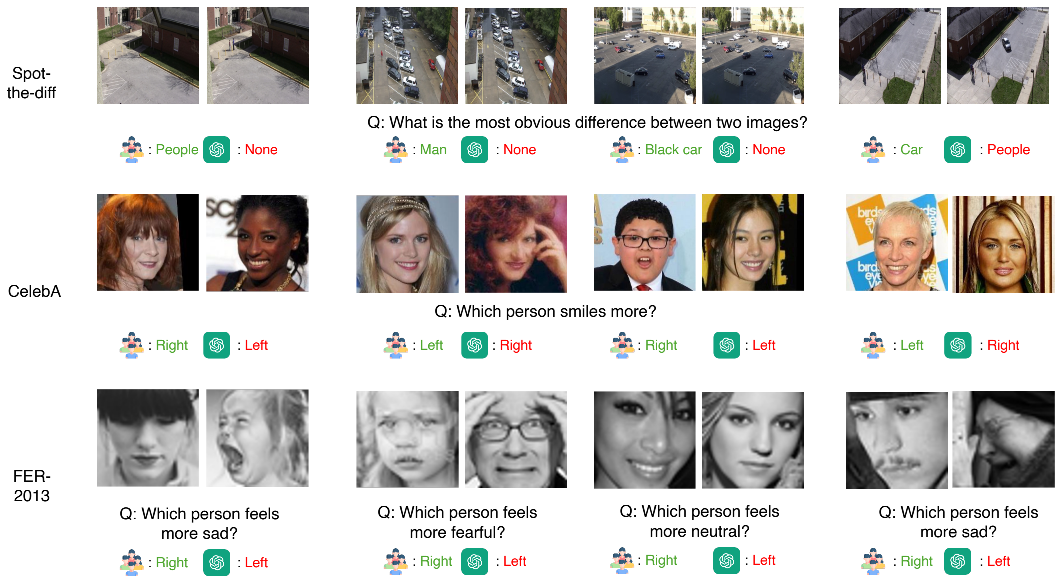

Figure 12: **Qualtiative examples on Spot-the-diff [6], CelebA [10], and FER-2013 [3].**

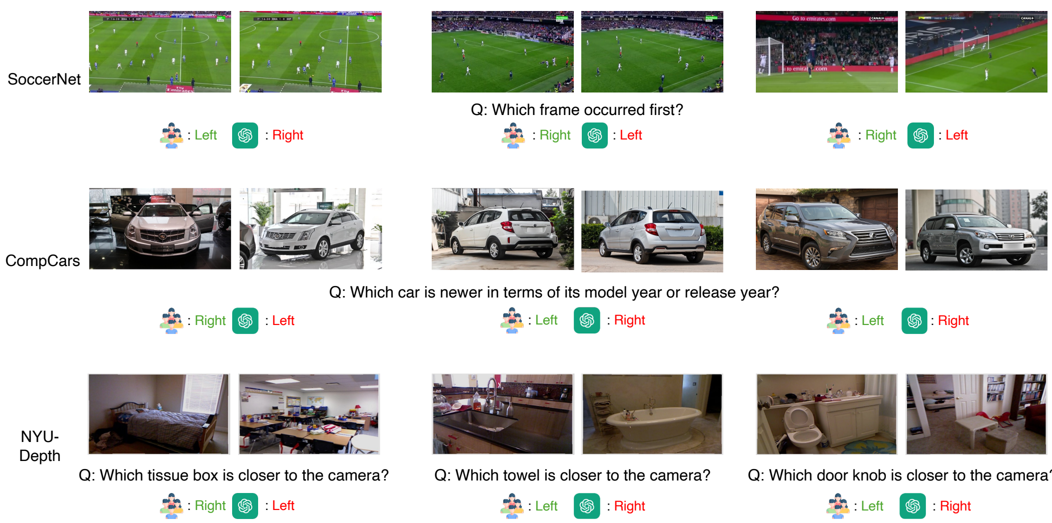

Figure 13: **Qualtiative examples on SoccerNet [2], CompCars [17], and NYU-Depth V2 [14].**

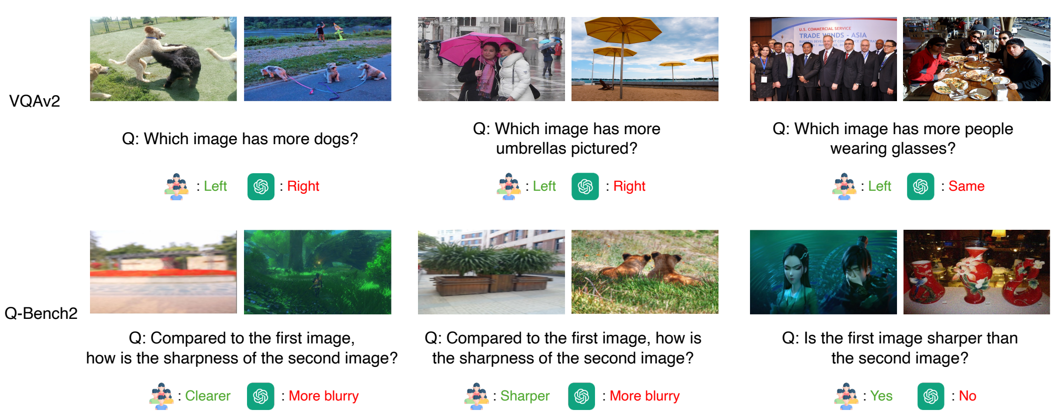

Figure 14: **Qualtiative examples on VQAv2 [4] and Q-Bench2 [19].**

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 6

[2] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *CVPR Workshops*, 2018. 2, 6, 7, 15

[3] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP*, 2013. 2, 6, 7, 15

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2, 6, 7, 15

[5] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 2, 3, 7, 14

[6] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *In EMNLP*, 2018. 2, 7, 15

[7] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020. 2, 5, 7, 14

[8] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *In CVPR*, 2024. 6

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *In NeurIPS*, 2024. 6

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *In ICCV*, 2015. 2, 6, 7, 15

[11] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, 2021. 2, 3, 7, 14

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[14] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5, 7, 15

[15] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011. 2, 3, 7, 14

[17] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015. 2, 6, 7, 15

[18] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *In NeurIPS*, 36, 2024. 1, 2, 7, 14

[19] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. A benchmark for multi-modal foundation models on low-level vision: from single images to pairs. *arXiv preprint arXiv:2402.07116*, 2024. 2, 6, 7, 15

[20] Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Transactions on Multimedia*, 23:3603–3617, 2020. 2, 4, 7, 14