Can Vision-Language Models Replace Human Annotators: A Case Study with CelebA Dataset

Haoming Lu
Picsart AI Research
Picsart Inc.
haoming.lu@picsart.com

Feifei Zhong
College of Computing
Georgia Institute of Technology
fzhong30@gatech.edu

Abstract

This study evaluates the capability of Vision-Language Models (VLMs) in image data annotation by comparing their performance on the CelebA dataset in terms of quality and cost-effectiveness against manual annotation. Annotations from the state-of-the-art LLaVA-NeXT model on 1000 CelebA images are in 79.5% agreement with the original human annotations. Incorporating re-annotations of disagreed cases into a majority vote boosts AI annotation consistency to 89.1% and even higher for more objective labels. Cost assessments demonstrate that AI annotation significantly reduces expenditures compared to traditional manual methods—representing less than 1% of the costs for manual annotation in the CelebA dataset. These findings support the potential of VLMs as a viable, cost-effective alternative for specific annotation tasks, reducing both financial burden and ethical concerns associated with large-scale manual data annotation. The AI annotations and re-annotations utilized in this study are available on GitHub.

1 Introduction

High-quality annotated data is recognized as a pivotal factor in the advancement of deep learning [6, 13]. Nevertheless, manual data annotation presents significant challenges in terms of cost and ethical considerations [5]. Recent evolution in large language models (LLMs) [21] has sparked enormous interest in their application to annotation and generation of text datasets [15, 3, 1, 18, 17]. Meanwhile, there are limited efforts [20, 7] exploring the capabilities of vision language models (VLMs) [19] in processing unlabeled image data. Previous studies have verified that VLMs can create various types of annotations on raw image data. However, comprehensive evaluations of their annotation quality and cost-effectiveness are essential to assess their potential to replace manual annotation.

In this paper, we evaluated the capability of AI-driven image data annotation by comparing the quality and costs between manual annotations and annotations generated by one of the SOTA VLMs (LLaVA-NeXT [10]) on the CelebA [11] dataset. Overall, the contributions of this work include:

- We confirmed that for specific image classification tasks, AI models can achieve performance comparable to human annotators, particularly excelling on more objective labels.
- Based on the similar quality and significantly lower costs, we argue that AI models already possess the potential to replace human annotation within certain scopes.

2 Background

CelebFaces Attributes Dataset (CelebA) [11] is a public face attributes dataset with over 200K celebrity images, each with 40 binary attribute annotations. Although the dataset was manually created, analysis [16] has pointed out errors and inconsistencies in the annotations.

LLaVA-NeXT [10] is an open-sourced SOTA multimodal model that enhances visual reasoning and OCR capabilities compared to LLaVA-1.5 [9], which served as the foundation of many comprehensive studies of data, model, and capabilities of large multimodal models (LMM).
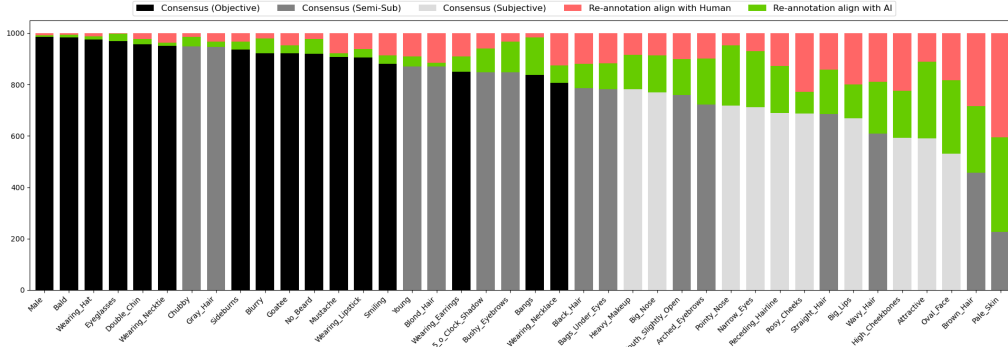
3   Experiments

Figure 1: The performance of AI-generated annotations on the 40 CelebA attributes: (1) The more objective the attribute, the higher the agreement between AI and human annotations; (2) In cases of discrepancy, re-annotation aligned more with AI in exactly 50% of the attributes.

For each of the 40 binary attributes in the CelebA dataset, we designed a question that required the model to respond with only **yes** or **no**. For instance, the question posed for the attribute **Eyeglasses** was: *Is this person wearing eyeglasses? Answer with only yes or no.* Annotations were generated on a randomly selected subset of 1000 images. Two reviewers re-annotated the attributes where AI and manual annotations differed. Inspired by [16], the attributes were divided into 3 groups based on their levels of objectivity. As illustrated in Figure 1, AI and the original human annotation reached **79.5%** overall consensus, with higher numbers on more objective attributes. Conducting a majority vote with the additional re-annotation further boosts the match rate between AI annotations and the final consensus to **89.1%**. Specifically, the re-annotation showed an equal preference for AI and human annotations, with each method favored more in exactly 20 attributes.

In our experiment, 40K labels for 1000 images were generated by a single NVIDIA RTX A6000 GPU at float16 precision using the HuggingFace implementation of the LLaVA-NeXT-8B model [12]. However, considering that in real life, the unit price of manual annotation decreases with scale, to fairly compare the costs of AI and human annotation, the cost estimation is based on constructing the entire CelebA dataset—generating 40 labels on each of the 200K images. The pricing for manual annotation is calculated based on publicly available quotes, whereas the cost of AI annotations is estimated from the average inference time per label and prices of renting GPU instances from Lambda [8]. As shown in Table 1, for image classification tasks like those in the CelebA dataset, AI annotation costs can be less than **1%** of human annotation costs.

	Human			AI	
Provider	Google [4]	AWS [14]	Scale AI [2]	A6000	A100
Cost (USD)	1100.00	530.00	2420.00	4.36	**4.01**

Table 1: The costs of manual and AI annotations (per 40K labels for 1000 images).

4   Conclusions and Limitations

In this study, we evaluated the differences in quality and costs between the CelebA dataset constructed using AI models and the one manually created. Given the comparable quality and substantial cost advantages, Vision-Language Models (VLMs) have shown the potential to replace one or more annotation sources in scenarios where multiple annotations are utilized to enhance accuracy. Limited by the experiment's scale, the results may have been influenced by the biases of the AI model and the annotators' subjectivity. Future research could focus on whether VLMs can effectively handle more complex tasks and whether better interactions with AI models can improve the quality of annotations.

References

[1] Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207, 2023.

[2] Scale AI. Pricing, 2024. https://scale.com/pricing, Accessed: 2024-09-10.

[3] Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media. In *World Conference on Explainable Artificial Intelligence*, pages 198–213. Springer, 2023.

[4] Google Cloud. Ai platform data labeling service pricing, 2024. https://cloud.google.com/ai-platform/data-labeling/pricing, Accessed: 2024-09-10.

[5] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

[6] Zeyad Emam, Andrew Kondrich, Sasha Harrison, Felix Lau, Yushi Wang, Aerin Kim, and Elliot Branson. On the state of data in computer vision: Human annotations remain indispensable for developing deep learning models. *arXiv preprint arXiv:2108.00114*, 2021.

[7] Bin Han, Yiwei Yang, Anat Caspi, and Bill Howe. Towards zero-shot annotation of the built environment with vision-language models (vision paper). *arXiv preprint arXiv:2408.00932*, 2024.

[8] Lambda Labs. Gpu cloud, 2024. https://lambdalabs.com/service/gpu-cloud, Accessed: 2024-09-10.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[12] LMMs-Lab. lmms-lab/llama3-llava-next-8b, 2024. https://huggingface.co/lmms-lab/llama3-llava-next-8b, Accessed: 2024-09-10.

[13] Christoffer Bøgelund Rasmussen, Kristian Kirk, and Thomas B Moeslund. The challenge of data annotation in deep learning—a case study on whole plant corn silage. *Sensors*, 22(4):1596, 2022.

[14] Amazon Web Services. Amazon sagemaker ground truth pricing, 2024. https://aws.amazon.com/sagemaker/groundtruth/pricing, Accessed: 2024-09-10.

[15] Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

[16] Haiyu Wu, Grace Bezold, Manuel Günther, Terrance Boult, Michael C King, and Kevin W Bowyer. Consistency and accuracy of celeba attribute values. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3258–3266, 2023.

[17] Jiannan Xiang, Zhengzhong Liu, Yucheng Zhou, Eric P Xing, and Zhiting Hu. Asdot: Any-shot data-to-text generation with pretrained language models. *arXiv preprint arXiv:2210.04325*, 2022.

[18] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.

[20] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, pages 159–175. Springer, 2022.

[21] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A Grouping of CelebA attributes

As shown in Table A, the 40 attributes in CelebA dataset are divided into three groups:

- Objective: These attributes are based on clear traits or easily identifiable elements, and are not open to personal interpretation. For example, whether the person is wearing glasses.
- Semi-Subjective: The attributes are based on ambiguous or less recognizable features that may be interpreted differently between observers. For instance, although youthfulness is a relatively objective attribute, people may have different perspectives on specific individuals.
- Subjective: These attributes are based on personal judgment and can vary significantly between observers. For example, whether a person is considered attractive entirely depends on the observer's preferences.

Group	Attributes
Objective	Male, Bald, Wearing_Hat, Eyeglasses, Double_Chin, Wearing_Necktie, Sideburns, Blurry, Goatee, No_Beard, Mustache, Wearing_Lipstick, Smiling, Wearing_Earrings, Bangs, Wearing_Necklace
Semi-Subjective	Chubby, Gray_Hair, Young, Blond_Hair, 5_o_Clock_Shadow, Bushy_Eyebrows, Black_Hair, Bags_Under_Eyes, Mouth_Slightly_Open, Arched_Eyebrows, Straight_Hair, Wavy_Hair, Brown_Hair, Pale_Skin
Subjective	Heavy_Makeup, Big_Nose, Pointy_Nose, Narrow_Eyes, Receding_Hairline, Rosy_Cheeks, Big_Lips, High_Cheekbones, Attractive, Oval_Face

Table 2: Grouping of the 40 CelebA attributes.

Despite the minor differences in the grouping used between our study and that of [16], the choice of grouping does not affect the conclusions.

B Detailed comparison of AI annotation, original human annotation, and re-annotation

Table 3 contains the detailed information that is illustrated in Figure 1. Column **AI=Human** is the number of images where the AI annotation is identical to the original human annotation; column **Re=AI** and **Re=Human** indicate the number of images where the re-annotation agrees with the AI annotation and the original human annotation respectively.

C Calculation of AI annotation costs

For inference with NVIDIA RTX A6000 GPU, the cost to annotate 1000 samples is estimated by

$$0.49 \times 40 \times 1000 \times \frac{1}{3600} \times 0.8 = 4.36$$

where 0.49 is the average inference time (seconds) per label, 40 and 1000 are the number of attributes per image and number of images, and 0.8 is the price per hour according to Lambda [8].

Similarly, for A100, the cost is

$$0.28 \times 40 \times 1000 \times \frac{1}{3600} \times 1.29 = 4.01$$

4

Attribute	AI=Human	Re=AI	Re=Human
Male	986	8	6
Bald	983	7	10
Wearing_Hat	976	12	12
Eyeglasses	970	2	28
Double_Chin	957	23	20
Wearing_Necktie	950	37	13
Chubby	949	14	37
Gray_Hair	947	34	19
Sideburns	937	34	29
Blurry	922	21	57
Goatee	922	47	31
No_Beard	919	22	59
Mustache	907	79	14
Wearing_Lipstick	906	61	33
Smiling	881	87	32
Young	871	91	38
Blond_Hair	870	116	14
Wearing_Earrings	850	90	60
5_o_Clock_Shadow	848	60	92
Bushy_Eyebrows	847	32	121
Bangs	838	17	145
Wearing_Necklace	807	125	68
Black_Hair	785	119	96
Bags_Under_Eyes	782	118	100
Heavy_Makeup	782	84	134
Big_Nose	770	87	143
Mouth_Slightly_Open	759	100	141
Arched_Eyebrows	722	98	180
Pointy_Nose	717	48	235
Narrow_Eyes	712	70	218
Receding_Hairline	689	128	183
Rosy_Cheeks	687	228	85
Straight_Hair	686	143	171
Big_Lips	669	199	132
Wavy_Hair	608	190	202
High_Cheekbones	592	224	184
Attractive	591	111	298
Oval_Face	531	183	286
Brown_Hair	456	284	260
Pale_Skin	226	406	368

Table 3: Detailed data underlying Figure 1.

D   Confidence intervals of the agreement percentages

Below are the agreement percentages between AI annotation and human re-annotation / majority votes by different label categories (subject, semi-subjective, objective).

Category	90%	95%	99%
Objective	(0.8960, 0.9428)	(0.8910, 0.9479)	(0.8800, 0.9588)
Semi-Subjective	(0.6457, 0.8338)	(0.6250, 0.8545)	(0.5797, 0.8997)
Subjective	(0.6272, 0.7208)	(0.6163, 0.7317)	(0.5910, 0.7570)
Overall	(0.7514, 0.8389)	(0.7427, 0.8477)	(0.7249, 0.8655)

Table 4: Confidence intervals of agreement between AI and human re-annotation by label categories.

Category	90%	95%	99%
Objective	(0.9463, 0.9766)	(0.9430, 0.9799)	(0.9360, 0.9869)
Semi-Subjective	(0.8214, 0.9159)	(0.8110, 0.9263)	(0.7883, 0.9490)
Subjective	(0.7699, 0.8505)	(0.7605, 0.8599)	(0.7387, 0.8817)
Overall	(0.8662, 0.9161)	(0.8611, 0.9211)	(0.8510, 0.9312)

Table 5: Confidence intervals of agreement between AI and majority votes by label categories.