

# Neurosymbolic Tag-Based Annotation for Interpretable Avatar Creation

Minghao Liu<sup>1,4,\*</sup> Zeyu Cheng<sup>2</sup> Shen Sang<sup>3</sup> Jing Liu<sup>3</sup> James Davis<sup>1,\*</sup>

<sup>1</sup>UC Santa Cruz <sup>2</sup>SJSU <sup>3</sup>ByteDance <sup>4</sup>Accenture

\*{mliu40,davisje}@ucsc.edu

**Editors:** Leilani H. Gilpin, Eleonora Giunchiglia, Pascal Hitzler, and Emile van Krieken

## Abstract

Avatar creation from human images presents challenges for direct neural approaches, which suffer from inconsistent predictions and poor interpretability due to the large parameter space with hundreds of ambiguous options. We propose a neurosymbolic tag-based annotation method that combines neural perceptual learning with symbolic semantic reasoning. Instead of directly predicting avatar parameters, our approach uses a neural network to predict semantic tags (hair length, curliness, direction) as an intermediate symbolic representation, then applies symbolic search algorithms to match optimal avatar assets. This neurosymbolic design produces higher annotator agreements (96.7% vs 31.0% for direct annotation), enables more consistent model predictions, and provides interpretable avatar selection with ranked alternatives. The tag-based system generalizes easily across rendering systems, requiring only new asset annotation while reusing human image tags. Experimental results demonstrate superior convergence, consistency, and visual quality compared to direct prediction methods, showing how neurosymbolic approaches can improve trustworthiness and interpretability in creative AI applications.

**Keywords:** Neurosymbolic AI, tag-based annotation, avatar creation

## 1. Introduction

Well-designed avatar creation tools like Bitmoji [Bitmoji](#), Google Cartoonset [Cloe et al. \(2022\)](#), and Metahuman [MetaHuman](#) provide expressive tools for users to create digital figures based on themselves. However, customizing the ideal avatar involves laborious selection and adjustment of parameters. Such a process consumes a significant amount of time from an average user without necessarily resulting in their ideal design. Training a learning-based algorithm for avatar auto-creation is needed.

Traditional neural approaches attempt to directly map human photographs to avatar parameters through end-to-end learning. However, this direct neural mapping suffers from fundamental limitations: the large parameter space with hundreds of ambiguous options leads to inconsistent predictions and lacks interpretability. Supervised learning requires the collection of pairwise training data, where annotators manually create corresponding avatars by selecting the best assets. Unfortunately, there are inherent issues with this *direct* annotation method. During the creation process, some parameters such as hairstyle include hundreds of options with only minor differences. It is almost impossible for the annotators to consistently select a single optimal choice, resulting in low agreement with other annotators. When collected in this way, the dataset has high label noise, and majority vote aggregation does little to help.

Instead, we propose a *neurosymbolic tag-based* annotation method for avatar creation that combines neural perceptual learning with symbolic semantic reasoning. Our approach



Figure 1: Avatar hairstyle prediction is hard due to hundreds of ambiguous options. Our neurosymbolic tag-based approach combines neural perception with symbolic reasoning, resulting in better labels (Sec 4.1), predictions (Sec 4.2), and generalizability (Sec 4.3).

introduces a semantically meaningful set of tags as an intermediate symbolic representation which applies to both photographs and avatar renderings. This neurosymbolic design bridges the gap between neural image understanding and symbolic avatar selection, providing interpretability while maintaining learning efficiency.

In our framework, annotators label both training photographs and stylized assets using semantic tags. For example, annotating *hair length*, *hair curly level*, and *hair direction* instead of simply *finding the best option out of two hundred hairstyles*. This symbolic representation provides better instructions and encourages annotators to search for more detailed semantic features when labeling. The proposed tag-based annotation results in higher annotator agreements due to the interpretable nature of the semantic categories.

Given a photograph, our neural component predicts semantic tags, which do not directly provide an answer for which asset is the best match. We then employ a symbolic search algorithm to evaluate the similarity between predicted image tags and the tags of each possible asset. The asset with maximum similarity is selected as the best match.

To evaluate how this neurosymbolic approach affects the final system, we compared supervised learning models trained on tag-based labels with models trained on direct labels. Experimental results show that models trained with the tag-based system produce better and more consistent predictions. Example hairstyles predicted by the neurosymbolic tag-based system are shown in Fig. 1.

Finally, we demonstrate that the neurosymbolic tag-based system enhances generalizability across different avatar rendering systems. When shifting to a new rendering system, direct neural approaches require completely new training sets of human-avatar pairs. In contrast, our symbolic tag representations of human images can be reused, and only the relatively small set of new avatar assets requires re-annotation. In a typical system, assets number in the hundreds while training photographs number in the tens of thousands, making this advantage significant.

This paper contributes a neurosymbolic method for avatar creation using *tag-based* annotation that demonstrates how hybrid neural-symbolic architectures can improve both trustworthiness and interpretability in creative AI applications. The advantages of this method include:

- Cleaner labels with higher annotator agreement via interpretable semantic tags
- Better and more consistent predictions from the neural-symbolic model
- Lower cost generalization to new rendering systems via reusable symbolic representations

## 2. Related work

**Image Stylization:** Creating a virtual character from an input human portrait image needs to overcome the domain gaps between the real world and the target styles. Gatys et al. matched feature information from CNN models to achieve style transfer [Gatys et al. \(2016\)](#). Cycle consistency loss was used for image-to-image transfer with paired-wise data supervision and with self-supervision [Isola et al. \(2017\)](#); [Zhu et al. \(2017\)](#); [Park et al. \(2020\)](#). Recently, the development of GAN inversion methods results in excellent image decomposition and high-quality reconstructions, which has been applied to image stylization [Richardson et al. \(2021\)](#); [Tov et al. \(2021\)](#); [Song et al. \(2021\)](#); [Cao et al. \(2018\)](#); [Zhu et al. \(2021\)](#). However, all of these methods focused on creating high-quality images in pixel space, as opposed to selecting assets in an avatar rendering system.

**Avatar creation using non-semantic parameters:** Creating avatars in parameter spaces without semantic meaning has been well-studied for many years. Extremely high quality methods for photorealistic avatars using stereo vision and single input images exist, with multiple good survey papers [Beeler et al. \(2010\)](#); [Yang et al. \(2020\)](#); [Blanz and Vetter \(1999\)](#); [Peng et al. \(2017\)](#); [Deng et al. \(2019\)](#); [Xu et al. \(2020\)](#); [Chen and Kim \(2021\)](#); [Egger et al. \(2020\)](#); [Zollhöfer et al. \(2018\)](#).

Stylized avatar systems also exist. Some methods utilize sketches as the prior condition for generation [Han et al. \(2017, 2018\)](#). Other methods are guided by position and landmarks, extracting human facial features used to deform textures and meshes [Wu et al. \(2018\)](#); [Cai et al. \(2021\)](#); [Lewiner et al. \(2011\)](#); [Vieira et al. \(2013\)](#). Recently a conditional GAN has been applied in the generation process [Li et al. \(2021\)](#); [Ye et al. \(2021\)](#). However, these methods all utilized parameters without semantic meaning, making them inapplicable to avatar systems designed to provide user level customization of asset choice.

**Avatar creation using semantic parameters:** To provide tools for customization of avatar creation, excellent rendering tools like Bitmoji, Metahuman, and Google Cartoon Set were created [Bitmoji](#); [MetaHuman](#); [Cloe et al. \(2022\)](#). These rendering systems provide explicit semantic meanings to each parameter and focus primarily on manual user creation.

Avatar prediction has been explored using self-supervised methods to avoid the difficulty of manual labeling. When the avatar is semi-photorealistic, F2P utilizes neural imitators to mimic the behaviors of the rendering system, improving efficiency and applying textures for more photorealistic visual quality [Shi et al. \(2019, 2020\)](#); [Lin et al. \(2021\)](#). In the stylized domain, AgileAvatar introduced a domain transfer module to the avatar creation pipeline [Sang et al. \(2022\)](#). However, these self-supervised methods rely heavily on carefully tuning each style. We provide a comparison to these methods in our results section.

**Human face datasets:** Training neural engines require the collection of human face datasets. FFHQ provides a collection of high-quality human face images without annotation [Karras et al. \(2019\)](#). CelebA and MAAD datasets include some basic tags of facial attributes [Liu et al. \(2015\)](#); [Terhörst et al. \(2021, 2019\)](#). FairFace includes ethnic tags and provides a racially balanced set [Karkkainen and Joo \(2021\)](#).

Hairstyle specific datasets also exist. Figaro-1k provides a limited set of samples, Hairstyle-30k treats the task as an end-to-end classification task, while K-hairstyle focuses on Korean hairstyles [Svanera et al. \(2016\)](#); [Yin et al. \(2017\)](#); [Kim et al. \(2021\)](#). None of these datasets has labels matching the specific avatar rendering systems we use in our work. We make use of the FairFace dataset for photographs of human faces.

**Symbolic Reasoning in Computer Vision:** The use of symbolic intermediate representations has shown significant promise in computer vision tasks requiring both perception and reasoning. Scene graph generation methods convert visual scenes into symbolic relationship graphs that enable structured reasoning [Xu et al. \(2017\)](#); [Johnson et al. \(2015\)](#). Visual question answering systems employ symbolic program synthesis to break down com-

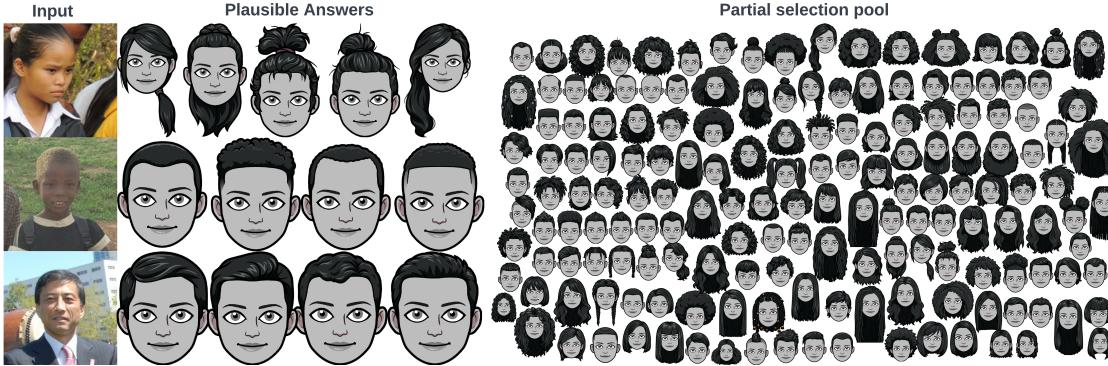


Figure 2: **Direct annotation challenges:** Hundreds of similar hairstyle options create ambiguity, with multiple plausible answers for each input image and no perfect matches, leading to poor annotator agreement.

plex queries into interpretable reasoning steps Johnson et al. (2017); Yi et al. (2018). Concept bottleneck models demonstrate how forcing models to predict human-interpretable concepts as intermediate representations improves both interpretability and performance Koh et al. (2020). Similarly, our semantic tag representation serves as an interpretable bottleneck that bridges neural image understanding and symbolic avatar selection, enabling transparent reasoning over hairstyle attributes.

### 3. Method

Direct annotation for avatar creation and the challenges it introduces are discussed in Sec. 3.1. Our neurosymbolic tag-based annotation system is introduced in Sec. 3.2. The symbolic search algorithm which relates tags to specific assets is discussed in Sec. 3.3. The neural vision backbone and training approach is provided in Sec. 3.4.

#### 3.1. Direct annotation challenges

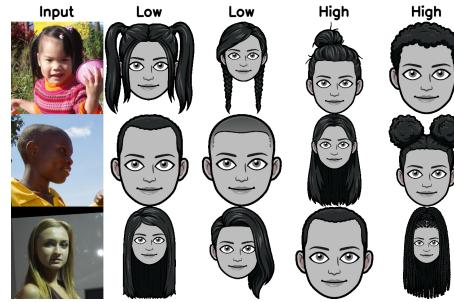
Customizing stylized avatars using rendering systems like Bitmoji requires tuning numerous parameters, some with hundreds of options. We focus on hairstyle prediction as it has the most variations, high visual impact, and significant ambiguity between similar assets. Fig. 2 shows sample Bitmoji hairstyles with the default face. Direct annotation asks annotators to select the best-matching asset from 200 options for each input photograph. However, this creates fundamental challenges: options often have subtle differences, no perfect matches exist, and multiple hairstyles may be plausible for a single input (examples shown left side of figure). This “no single best answer” phenomenon results in high label noise and low annotator agreement, making majority vote aggregation ineffective. These limitations of direct mapping motivate our neurosymbolic approach.

#### 3.2. Neurosymbolic tag-based annotation

In this paper, we propose a neurosymbolic tag-based annotation approach that introduces semantic tags as an intermediate symbolic representation. The goal is to map both human face images and avatar hairstyles to a semantically meaningful tag space. We defined our tags as in Fig. 3(a). Instead of providing the annotators with a massive number of options, we specifically ask them to annotate symbolic tag attributes from each region, for example *Hair direction on the top of the head*, or *Hair curliness level on the side of the head*. This

Region	Annotation Tags		Distance calculation	
	Attributes	# Options	Weight	Type
Top and front	Length	6	2.25	Continuous
	Direction	8	2	Discrete
	Curly level	4	1	Continuous
On the side	Length	5	2.25	Continuous
	Curly level	4	1	Continuous
Braid	Yes / No	2	5	Discrete
	Count	4	2	Discrete
	Position	3	1	Discrete
	Type	5	1	Discrete

(a) Semantic tag design



(b) Symbolic search visualization

Figure 3: **Neurosymbolic tag-based annotation system:** (a) **Semantic tag design** breaks down hairstyles into interpretable attributes (length, direction, curliness, braids) with defined options and weights, serving as intermediate representation between neural image understanding and avatar selection. (b) **Symbolic search visualization** shows how distance scores from our search algorithm measure semantic similarity—low distances produce visually similar matches while high distances indicate poor matches.

symbolic representation bridges the gap between neural image understanding and avatar selection. Detailed descriptions of each tag are included in the supplemental material.

Designing the appropriate semantic tags to describe the hairstyle requires domain knowledge. Each tag requires a clear definition. For example, *short hair* and *medium-short hair* is insufficient description for consistent labeling. We use an iterative design process to arrive at our final tag definitions. The researchers first designed tags to describe the hairstyles by simply looking at a set of human images and avatars. An annotator tagged all the avatar hairstyles using this tag design. A different annotator tagged a set of human images. Using the tags, the best matched avatar to each photo is retrieved. The researchers then evaluate the agreement between annotators, and the expressibility of the tag design, to make modifications to the set of tags. The process was repeated until tag design was considered sufficient.

After arriving at a tag design, we perform the complete run of data annotation. Note that our tag design pipeline allows researchers to focus on iteratively improving their symbolic representations while not requiring them to work as annotators. By going through such a design process, researchers verify their designs, so that higher agreement between researchers and annotators is achieved.

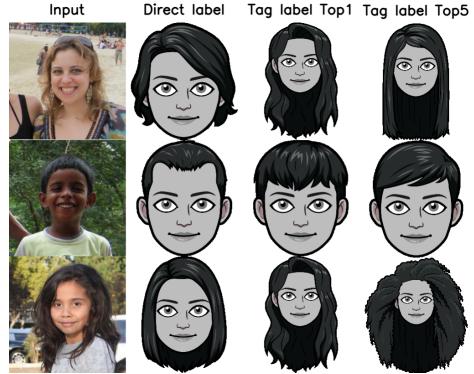
### 3.3. Symbolic search algorithm

Our designed tag system has 460,800 permutations, making it impossible to design a hairstyle for each permutation. This implies that for many human images there is no perfect hairstyle match. To address this issue, we employ symbolic reasoning to search through all existing hairstyles, computing semantic tag similarity. The overall distance of a particular asset is computed as a weighted sum of individual tag distances. The weight of each attribute is listed in Fig. 3(a). To measure the tag distance for each attribute, we used L1 loss for continuous variables, and zero-one loss for discrete variables.

The distance score from the symbolic search provides ranking information for all the hairstyles, while direct annotation only provides the Top-1 result. Fig. 3(b) shows visual samples of low and high-distance pairs. Note that the low-distance hairstyles have better visual similarity with the inputs, while the high-distance samples are visually dissimilar. This symbolic reasoning component enables interpretable ranking and alternative suggestions.

		Direct annotation	Tag-based annotation	
Chance Agreement exists	Tag level	NA	<b>96.7%</b>	
	Final Top-1	31.0%	<b>52.0%</b>	
	Final Top-2	NA	<b>67.2%</b>	
	Final Top-3	NA	<b>73.5%</b>	
	Final Top-4	NA	<b>80.3%</b>	
Time	Skilled annotators	25.1s	<b>23.6s</b>	
	Random Turker	<b>48.4s</b>	112.6s	
User study	Matching	89.4%	<b>92.7%</b>	
	Preference	306 : 306		

(a) Numerical comparisons



(b) Visual comparison

Figure 4: **Annotation method comparisons:** (a) **Numerical results** show our tag-based approach achieves dramatically higher annotator agreement (96.7% vs 31.0%) and enables ranked Top-K alternatives with 80.3% agreement for Top-4 aggregation, while direct annotation provides only single predictions. Time costs remain similar for skilled annotators. (b) **Visual comparison** shows both methods produce plausible results, but our neurosymbolic approach provides additional ranked alternatives through symbolic search.

### 3.4. Neural network training

We trained our neural component in a supervised manner to predict semantic tags from human images. To extract feature information from the image, we used the open-source pre-trained ResNet-50 He et al. (2016) from the PyTorch Paszke et al. (2017) library as our vision backbone and the initial training checkpoint. During training, we used L2 loss for continuous tag variables and cross-entropy loss for discrete tag classifications. The neural network learns to map from image pixels to semantic tag representations, which are then processed by the symbolic search component.

## 4. Results and Experiments

In this section, we demonstrate the advantages of our neurosymbolic tag-based annotation approach with experimental results. Sec. 4.1 shows the advantage at the annotation level, Sec. 4.2 shows the advantage on neural model convergence and consistent predictions, and Sec. 4.3 shows that a neurosymbolic tag-based system can easily be compatible with new rendering systems.

### 4.1. Annotation Quality

**Annotator Agreement:** Label noise is a common problem in supervised learning models Wei et al. (2022). Collecting multiple copies of annotation for aggregation is often required to create a high-quality dataset. Using a majority vote is the most common way to aggregate labels and reduce label noise. However, given a large number of hairstyle options, there might exist multiple plausible answers, or alternatively, no ideal match and only partially correct answers. These situations cause low agreements between annotators. Fig. 4(a) provides evidence of the severity of the problem when 3 annotators provide independent labels for each target. Agreement exists between annotators in only 31.0% of cases when using direct annotation. In the majority of cases all three annotators provide different answers. On the other hand, annotator agreement on semantic tags exists in 96.7% of cases

when using our neurosymbolic tag annotations. In order to fairly compare against direct annotation, we compute agreement on the Top-1 hairstyle chosen by each set of tag labels. Because multiple similar hairstyles exist, agreement is lower than on raw tag prediction at only 52.0%, but this is substantially better (+21.0%) than the agreement when labelers provide direct annotations. Unlike direct annotation, our neurosymbolic tag-based annotation provides additional plausible answers which can be used to improve agreement and enable aggregation. Aggregating over the Top-4 matches from each of 3 labelers, reaches an agreement level of 80.39% (+49.37%). We conclude that neurosymbolic tag-based labels are substantially more valuable in creating a "clean" set of labels.

**Annotation cost:** The total cost of annotation is a function of labeling time for a individual image, and the total number of labels required. The cost of individual annotation is measured in time as shown in Fig. 4(a). We asked annotators from two different skill levels to annotate images and recorded annotation time. Our experiment shows well-trained professional annotators need marginally less time to annotate a face when using the neurosymbolic tag-based system. In contrast, untrained workers obtained from Amazon Mechanical Turk need more time when using tag-based labels. In either case, the differences in individual labeling time are bounded and not the most significant factor.

Direct annotation requires a completely new set of labels each time a change is made to the set of available hairstyles, with associated retraining of the prediction model given the new labels. In contrast, our neurosymbolic tag-based systems only need to label the new hairstyles, with no new tag labels on the much larger set of training images, and no retraining of the neural tag predictor. Since most avatar rendering systems will be updated with new artistic assets occasionally, we conclude that there is a substantial savings on label cost using neurosymbolic tag-based annotation.

**Visual Quality comparison:** To compare the visual quality between direct and neurosymbolic tag-based annotation, we conduct two user studies through Amazon Mechanical Turk [Amazon Mechanical Turk](#): *Matching* and *Preference*. In the *Matching* task, we evaluate whether visual similarity is sufficiently close that evaluators can tell which avatar goes with which human. A human image is shown with the corresponding avatar hairstyle and three random distractor hairstyles. The evaluator is required to match the human image and avatar image and is scored as a correct match if the evaluator correctly picks the original annotation. A high matching score indicates the avatar represents the human well. A total of 1,224 judgments were collected. As the result shows in Fig. 4(a), our proposed neurosymbolic tag-based annotation preserves user identity marginally better compared to direct annotation. In the *Preference* task, avatar results from both methods are presented for comparison with the human image. The evaluators were asked to provide their preferences by choosing one of the results or indifferent, a total of 612 judgments were collected. The results showed a precise split of 306 judgements for each method. The combined results of both studies indicate that our semantic tags are sufficiently expressive to act as a replacement for direct annotation in terms of visual quality.

A visual comparisons of the two annotations methods is provided in Fig. 4(b). Both annotation methods result in plausible answers. In addition to the Top-1 match, our neurosymbolic tag-based system can provide interpretable ranking information for all other hairstyles through symbolic search. The figure also shows Top-5 results from the tag-based annotations. The visual quality of these results is reduced, but they remain plausible.

## 4.2. Neural Model Prediction Quality

Neurosymbolic tag-based labels result in better neural models when used for selecting avatar assets, in terms of both visual quality and consistency.

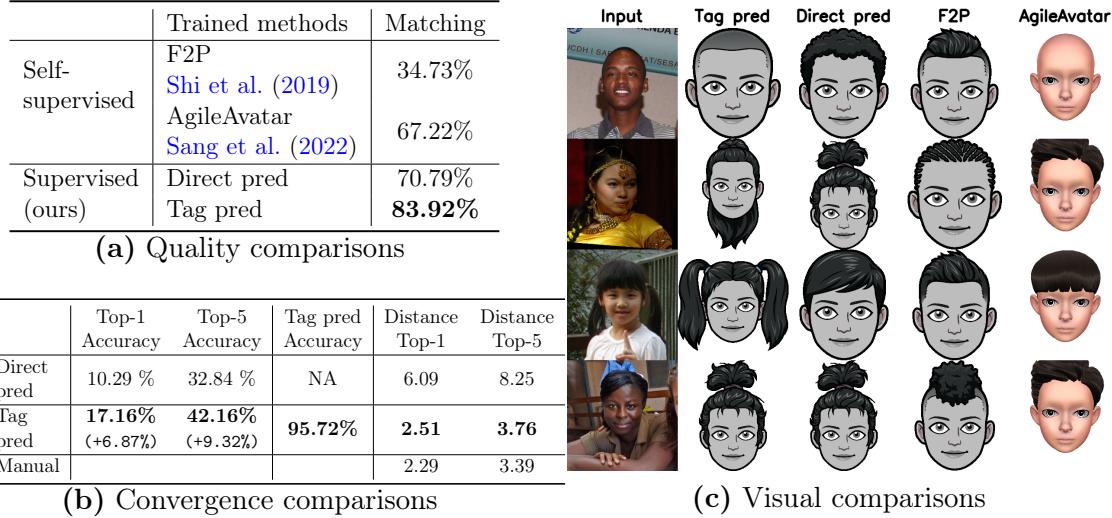


Figure 5: **Model prediction comparisons:** (a) **Quality results** show our neurosymbolic Tag pred achieves best matching performance (83.92%), outperforming direct supervised prediction (70.79%) and self-supervised baselines. (b) **Convergence metrics** demonstrate tag-based training improves accuracy (+6.87% Top-1, +9.32% Top-5) and reduces semantic distance scores. (c) **Visual results** show supervised methods outperform self-supervised approaches, with our neurosymbolic method capturing detailed semantic features through intermediate tag representation.

#### 4.2.1. BASELINE METHODS AND DATASET

We compare several methods to understand the effects of neurosymbolic tag-based annotation. We choose two state-of-the-art self-supervised baselines: F2P [Shi et al. \(2019\)](#), an optimization-based method for realistic game character creation, and AgileAvatar [Sang et al. \(2022\)](#), the SOTA learning-based method for stylized avatar creation. For supervised baselines, we compare Direct pred (treating the task as classification, predicting the best hairstyle with direct annotation targets) and Tag pred (our neurosymbolic approach using semantic tags and symbolic search as described in Sec. 3.3). Both supervised baselines use identical vision backbones with similar loss functions.

We used human face images from the FairFace [Karkkainen and Joo \(2021\)](#) dataset, which is racially balanced and includes blurry images, requiring model robustness on diverse and lower-quality inputs. We collected neurosymbolic tag-based annotations for 17k images using professional annotators: 14.5k for training, 2.8k for testing, and 204 as a holdout set for human evaluation studies.

To avoid bias from annotation quality differences, we create direct labels using tag-based labels via symbolic search, treating these hairstyles as training targets for Direct Pred. Thus both supervised methods share identical Top-1 training targets.

#### 4.2.2. BETTER PREDICTION QUALITY

**Visual comparisons for models:** Fig. 5(c) compares all four methods. F2P [Shi et al. \(2019\)](#), designed for realistic avatars, fails frequently on stylized avatars. AgileAvatar [Sang et al. \(2022\)](#) uses a stylization module to overcome domain gaps, significantly improving over F2P but remaining inferior to supervised methods. Comparing Direct pred and Tag

pred, tag-based training helps the neural component capture detailed semantic features like double pony-tails and hair curliness.

**Numerical comparison for models:** User studies (Fig. 5(a)) use matching tests where evaluators identify corresponding avatars among distractors, measuring identity preservation quality. F2P performs poorly due to its photorealistic design. AgileAvatar reaches similar scores to Direct pred (only 3.57% lower), but our neurosymbolic approach performs best, preserving more user identity than all baselines.

#### 4.2.3. BETTER CONVERGENCE, MORE CONSISTENT PREDICTIONS

**Convergence:** Both supervised models target the same best-matching hairstyles—explicitly for direct prediction, implicitly through semantic tags for our approach. Fig. 5(b) shows tag prediction achieves 95.72% accuracy versus 10.29% for direct prediction, but this comparison is unfair due to different class numbers. For fair comparison, we evaluate final hairstyle selection among hundreds of options using Top-K accuracy. Our neurosymbolic training achieves better Top-1 (+6.87%) and Top-5 (+9.32%) accuracy.

The low absolute Top-1 accuracy (17.16%) reflects the challenge of hundreds of similar options with multiple potential correct answers rather than poor quality. Our symbolic search distance metric provides better quality assessment, measuring semantic similarity between human faces and avatar predictions using annotator-provided tags. Our Tag pred achieves lower average symbolic search distances for Top-1 (2.51) and Top-5 (3.76) predictions compared to Direct pred (6.09, 8.25). Manual distances provide lower bounds since perfect matches are impossible.

**Consistency:** We visualize the Top-5 predictions of both supervised learning methods in Fig. 6(a). While both methods produce plausible Top-1 answers, the Top-5 predictions from our neurosymbolic *Tag Pred* have better consistency compared to *Direct Pred*. The *Direct Pred* model treats each hairstyle as an independent class without considering their symbolic similarities to the human image. Our neurosymbolic *Tag pred*, on the other hand, trains the neural component to predict semantic features defined by human researchers, then uses symbolic reasoning to find consistent matches. Thus the model was encouraged to focus on the interpretable features that are important to human observers, resulting in consistent Top-K predictions. Notice for example that even when direct prediction correctly predicts a short hairstyle as the Top-1 result, the next best prediction might be long hair.

### 4.3. Generalizability

Annotating datasets requires substantial effort—in our case, 17k sets of semantic tags for human images and avatar hairstyles. Direct annotation requires completely new labels and model retraining for each rendering system. Our neurosymbolic approach significantly reduces this cost since semantic tags for human images are rendering-system independent. Only new avatar assets require tagging ( 200 hairstyles vs. 17k training images, <2% of original cost), while the neural tag prediction model remains valid without retraining.

To demonstrate generalizability, we collected tags for avatar samples from four diverse systems: Bitmoji [Bitmoji](#) (cartoon avatars with gender-neutral, gray-scaled default faces), Google Cartoonset [Cloe et al. \(2022\)](#) (cartoon dataset with random non-hairstyle attributes), MetaHuman [MetaHuman](#) (realistic avatars with gender selection based on Fair-Face tags), and NovelAI [NovelAI](#) (diffusion-based cartoon generation using artist-selected text prompts). We controlled only hairstyles across all systems.

Fig. 6(b) shows model-predicted results. Given semantic tag predictions from our neural component, symbolic search finds the closest matching hairstyle in each system based on

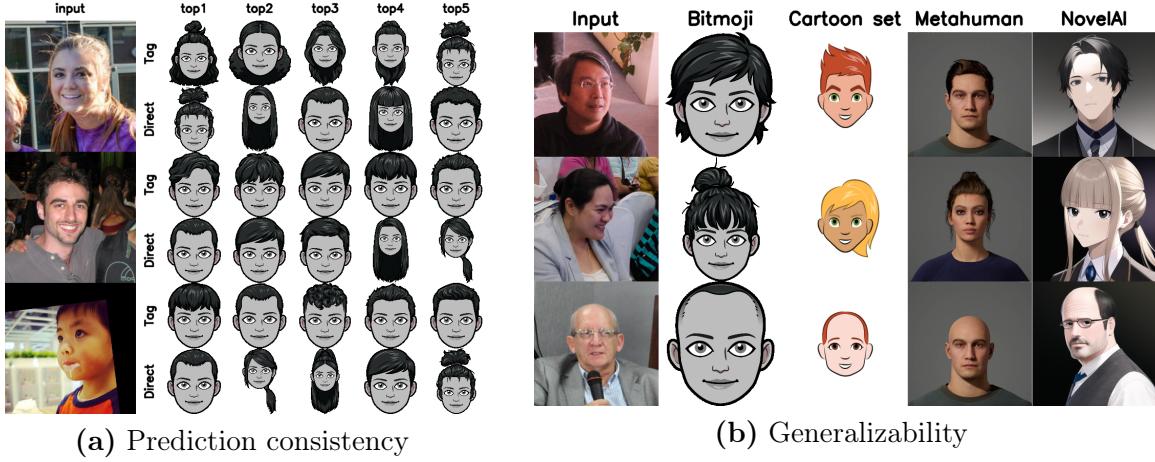


Figure 6: **Neurosymbolic system advantages:** (a) **Prediction consistency** shows our neurosymbolic approach produces more coherent Top-5 rankings than direct prediction, which can include contradictory results (e.g., both short and long hair). (b) **Cross-system generalizability** demonstrates that semantic tags learned on one avatar system transfer to other rendering systems without retraining, requiring only new asset annotation.

avatar tags. While not every system contains every hairstyle, the selected avatars provide good approximations of input photographs across diverse rendering approaches.

## 5. Limitations and conclusion

**Limitations:** To benefit from the neurosymbolic tag-based system, carefully designed semantic tags with clear definitions are required, demanding domain expertise and iterative refinement. The symbolic search algorithm is sensitive to neural tag prediction errors—while our neural component substantially outperforms direct prediction methods, incorrect semantic tags mislead the symbolic search component, highlighting the importance of robust neural training.

As with all avatar prediction methods, our work lacks universally accepted benchmarks. Common metrics (L2 loss, perceptual loss Johnson et al. (2016), Top-K accuracy) poorly represent user preferences. While we conducted user studies, evaluators may not reflect actual users, affecting evaluation of both neural and symbolic components.

**Conclusion:** We present a neurosymbolic tag-based annotation method for avatar creation that demonstrates how hybrid neural-symbolic architectures improve trustworthiness and interpretability in creative AI applications. Our approach combines neural perceptual learning with symbolic semantic reasoning through intermediate tag representations, achieving higher annotation quality (96.7% vs. 31.0% agreement), better model convergence and consistency, and easy generalization to new rendering systems with minimal cost (<2% of original annotation effort). Experimental results demonstrate superior performance over direct neural methods across annotation quality, model training, and system generalizability, contributing to understanding how neurosymbolic AI can enhance creative applications.

## References

- Amazon Mechanical Turk. Amazon mechanical turk. <https://www.mturk.com>. Accessed: 2022-11-10.
- Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9. ACM New York, NY, USA, 2010.
- Bitmoji. Bitmoji. <https://www.bitmoji.com/>. Accessed: 2022-11-10.
- Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.
- Hongrui Cai, Yudong Guo, Zhuang Peng, and Juyong Zhang. Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model. *Graphical Models*, 115:101103, 2021.
- Kaidi Cao, Jing Liao, and Lu Yuan. Carigans: Unpaired photo-to-caricature translation, 2018.
- Zhixiang Chen and Tae-Kyun Kim. Learning feature aggregation for deep 3d morphable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13164–13173, 2021.
- Forrester Cloe, Inbar Mosseri, Dilip Krishnan, Aaron Sarna, Aaron Maschinot, Bill Freeman, and Shiraz Fuman. Cartoon set, 2022. data retrieved from Google Machine Perception organization, <https://google.github.io/cartoonset/people.html>.
- Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.
- Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- Xiaoguang Han, Chang Gao, and Yizhou Yu. Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling. *ACM Transactions on graphics (TOG)*, 36(4):1–12, 2017.
- Xiaoguang Han, Kangcheng Hou, Dong Du, Yuda Qiu, Shuguang Cui, Kun Zhou, and Yizhou Yu. Caricatureshop: Personalized and photorealistic caricature sketching. *IEEE transactions on visualization and computer graphics*, 26(7):2349–2361, 2018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017.
- Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.
- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- Taewoo Kim, Chaeyeon Chung, Sunghyun Park, Gyojung Gu, Keonmin Nam, Wonzo Choe, Jaesung Lee, and Jaegul Choo. K-hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1299–1303. IEEE, 2021.
- Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348, 2020.
- Thomas Lewiner, Thales Vieira, Dimas Martínez, Adelailson Peixoto, Vinícius Mello, and Luiz Velho. Interactive 3d caricature from harmonic exaggeration. *Computers & Graphics*, 35(3): 586–595, 2011.
- Song Li, Songzhi Su, Juncong Lin, Guorong Cai, and Li Sun. Deep 3d caricature face generation with identity and structure consistency. *Neurocomputing*, 454:178–188, 2021.
- Jiangke Lin, Yi Yuan, and Zhengxia Zou. Meingame: Create a game character face from a single portrait. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 311–319, 2021.
- 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)*, December 2015.
- MetaHuman. Unreal engine metahuman. <https://www.unrealengine.com/en-US/metahuman>. Accessed: 2022-11-10.
- NovelAI. Novelai. <https://novelai.net/>. Accessed: 2022-11-10.
- Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017. URL <https://openreview.net/forum?id=BJJsrmfCZ>.

Weilong Peng, Zhiyong Feng, Chao Xu, and Yong Su. Parametric t-spline face morphable model for detailed fitting in shape subspace. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6139–6147, 2017.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Shen Sang, Tiancheng Zhi, Guoxian Song, Minghao Liu, Chunpong Lai, Jing Liu, Xiang Wen, James Davis, and Linjie Luo. Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. *ACM SIGGRAPH Asia 2022 Conference Proceedings*, 2022.

Tianyang Shi, Yi Yuan, Changjie Fan, Zhengxia Zou, Zhenwei Shi, and Yong Liu. Face-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 161–170, 2019.

Tianyang Shi, Zhengxia Zuo, Yi Yuan, and Changjie Fan. Fast and robust face-to-parameter translation for game character auto-creation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1733–1740, 2020.

Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

Michele Svanera, Umar Riaz Muhammad, Riccardo Leonardi, and Sergio Benini. Figaro, hair detection and segmentation in the wild. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 933–937. IEEE, 2016.

Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Ines Zelch, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, September 23-26, 2019*, pages 1–8. IEEE, 2019. doi: 10.1109/BTAS46853.2019.9185975. URL <https://doi.org/10.1109/BTAS46853.2019.9185975>.

Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maad-face: A massively annotated attribute dataset for face images. *IEEE Trans. Inf. Forensics Secur.*, 16:3942–3957, 2021. doi: 10.1109/TIFS.2021.3096120. URL <https://doi.org/10.1109/TIFS.2021.3096120>.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Roberto C Cavalcante Vieira, Creto A Vidal, and Joaquim Bento Cavalcante-Neto. Three-dimensional face caricaturing by anthropometric distortions. In *2013 XXVI Conference on Graphics, Patterns and Images*, pages 163–170. IEEE, 2013.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=TBWA6PLJZQm>.

Qianyi Wu, Juyong Zhang, Yu-Kun Lai, Jianmin Zheng, and Jianfei Cai. Alive caricature from 2d to 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7336–7345, 2018.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2020.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zipeng Ye, Mengfei Xia, Yanan Sun, Ran Yi, Minjing Yu, Juyong Zhang, Yu-Kun Lai, and Yong-Jin Liu. 3d-carigan: an end-to-end solution to 3d caricature generation from normal face photos. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in neural information processing systems*, pages 1031–1042, 2018.

Weidong Yin, Yanwei Fu, Yiqing Ma, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Learning to generate and edit hairstyles. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1627–1635, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks, 2021.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018.

**Appendix**

Here we include extra image samples for the figures in our paper:

- Sec. [A](#) provides mathematical formulation and detailed algorithms for our neurosymbolic approach
- Sec. [B](#) provides additional samples for Visualization of low and high distance samples (Fig3)
- Sec. [C](#) provides additional samples for Visual comparison of annotation results (Fig4)
- Sec. [D](#) provides additional samples for Visual comparisons of model predictions (Fig5)
- Sec. [E](#) provides additional samples for Prediction consistency (Fig6)
- Sec. [F](#) provides additional samples for Easily expandable to other systems (Fig7)
- Sec. [G](#) shows the annotator user interface

## Appendix A. Mathematical Formulation and Algorithm

### A.1. Problem Formulation

Let  $I$  be an input human face image and  $A = \{a_1, a_2, \dots, a_N\}$  be the set of available avatar hairstyles. Our goal is to find the optimal avatar  $a^* \in A$  that best matches the input image  $I$ .

#### A.1.1. SEMANTIC TAG SPACE

We define a semantic tag space  $T$  consisting of  $K$  attributes organized into regions  $R = \{r_1, r_2, \dots, r_M\}$ . Each attribute  $t_k$  can be either continuous or discrete:

$$T = \{t_1, t_2, \dots, t_K\} \text{ where } t_k \in \begin{cases} [0, 1] & \text{if continuous} \\ \{0, 1, 2, \dots, C_k\} & \text{if discrete} \end{cases}$$

The complete tag representation for an image or avatar is:

$$\mathbf{t} = [t_1, t_2, \dots, t_K]^T$$

In our implementation, we organize attributes into three main regions:

- **Top/Front region:** Hair length, direction, and curliness level
- **Side region:** Hair length and curliness level
- **Braid region:** Presence, count, position, and type

#### A.1.2. NEURAL TAG PREDICTION

A neural network  $f_\theta$  with parameters  $\theta$  maps from image space to tag space:

$$\mathbf{t}_I = f_\theta(I)$$

The network is trained using a composite loss function that handles both continuous and discrete attributes:

$$\mathcal{L}(\theta) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left[ \sum_{k \in \mathcal{C}} \|\hat{t}_{i,k} - t_{i,k}\|_2^2 + \sum_{k \in \mathcal{D}} \text{CE}(\hat{t}_{i,k}, t_{i,k}) \right]$$

where  $\mathcal{C}$  and  $\mathcal{D}$  are the sets of continuous and discrete attributes respectively,  $\text{CE}$  denotes cross-entropy loss, and  $\hat{t}_{i,k}$  represents the predicted tag value for attribute  $k$  of sample  $i$ .

#### A.1.3. SYMBOLIC SEARCH ALGORITHM

Given the predicted tags  $\mathbf{t}_I$  for an input image, we compute the semantic distance to each avatar  $a_j$  using a weighted combination of attribute-specific distances:

$$d(I, a_j) = \sum_{k=1}^K w_k \cdot d_k(\mathbf{t}_{I,k}, \mathbf{t}_{a_j,k})$$

where  $w_k$  is the importance weight for attribute  $k$ , and the attribute-specific distance function is:

$$d_k(\mathbf{t}_{I,k}, \mathbf{t}_{a_j,k}) = \begin{cases} |\mathbf{t}_{I,k} - \mathbf{t}_{a_j,k}| & \text{if continuous (L1 loss)} \\ \mathbb{I}[\mathbf{t}_{I,k} \neq \mathbf{t}_{a_j,k}] & \text{if discrete (0-1 loss)} \end{cases}$$

The optimal avatar is selected as:

$$a^* = \arg \min_{a_j \in A} d(I, a_j)$$

Additionally, we can rank all avatars by their distances to provide alternative suggestions:

$$\text{ranking}(I) = \text{sort}(\{(d(I, a_j), a_j)\}_{j=1}^N)$$

## A.2. Algorithms

---

**Input :** Input image  $I$ , Avatar set  $A = \{a_1, \dots, a_N\}$ , Neural network  $f_\theta$ , Avatar tags  $\{\mathbf{t}_{a_j}\}_{j=1}^N$

**Output:** Best matching avatar  $a^*$ , ranked alternatives, predicted tags  $\mathbf{t}_I$

```

// Neural Tag Prediction
 $\mathbf{t}_I \leftarrow f_\theta(I)$  ; // Predict semantic tags from input image

// Symbolic Search Phase
Initialize distance list  $D \leftarrow \emptyset$ 
for  $j \leftarrow 1$  to  $N$  do
     $d_j \leftarrow 0$  ; // Initialize distance for avatar  $a_j$ 
    for  $k \leftarrow 1$  to  $K$  do
        if attribute  $k$  is continuous then
            |  $d_k \leftarrow |\mathbf{t}_{I,k} - \mathbf{t}_{a_j,k}|$  ; // L1 distance
        else
            |  $d_k \leftarrow \mathbb{I}[\mathbf{t}_{I,k} \neq \mathbf{t}_{a_j,k}]$  ; // 0-1 distance
        end
         $d_j \leftarrow d_j + w_k \cdot d_k$  ; // Weighted accumulation
    end
     $D \leftarrow D \cup \{(d_j, a_j)\}$  ; // Store distance-avatar pair
end

// Ranking and Selection
Sort  $D$  by distance in ascending order  $a^* \leftarrow \text{avatar of } \arg \min_{(d,a) \in D} d$  ; // Best match
alternatives  $\leftarrow$  top-5 avatars from sorted  $D$ 

return  $a^*, \text{alternatives}, \mathbf{t}_I$ 

```

**Algorithm 1:** Neurosymbolic Tag-Based Avatar Selection

---

---

**Input** : Training set  $\mathcal{D} = \{(I_i, \mathbf{t}_i)\}_{i=1}^{N_{\text{train}}}$ , Learning rate  $\alpha$ , Number of epochs  $E$   
**Output:** Trained network parameters  $\theta^*$

Initialize network parameters  $\theta$  randomly

```

for epoch  $\leftarrow 1$  to E do
    for each mini-batch  $\mathcal{B} \subseteq \mathcal{D}$  do
         $\mathcal{L}_{\text{batch}} \leftarrow 0$  ;                                // Initialize batch loss
        for  $(I_i, \mathbf{t}_i) \in \mathcal{B}$  do
             $\hat{\mathbf{t}}_i \leftarrow f_\theta(I_i)$  ;                      // Forward pass
            for  $k \leftarrow 1$  to K do
                if attribute  $k$  is continuous then
                     $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \|\hat{t}_{i,k} - t_{i,k}\|_2^2$  ;      // MSE loss
                else
                     $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \text{CE}(\hat{t}_{i,k}, t_{i,k})$  ;          // Cross-entropy loss
                end
            end
        end
         $\mathcal{L}_{\text{batch}} \leftarrow \frac{\mathcal{L}_{\text{batch}}}{|\mathcal{B}|}$  ;                                // Average over batch
         $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{batch}}$  ;                         // Gradient descent update
    end
end
return  $\theta$ 

```

**Algorithm 2:** Neural Network Training for Tag Prediction

---

### A.3. Complexity Analysis

The computational complexity of our approach consists of two main components:

**Neural Tag Prediction:** The forward pass through the neural network has complexity  $O(P)$ , where  $P$  is the number of parameters in the network (typically ResNet-50 with  $\approx 25M$  parameters).

**Symbolic Search:** Computing distances to all avatars has complexity  $O(N \cdot K)$ , where  $N$  is the number of avatar hairstyles (typically  $\approx 200$ ) and  $K$  is the number of semantic attributes (in our case,  $K = 9$ ). Sorting for ranking adds  $O(N \log N)$ .

The total inference complexity is  $O(P + N \cdot K + N \log N)$ , which is dominated by the neural network forward pass in practice.

### A.4. Tag Design Specifications

Based on our iterative design process described in Section 3.2, the final semantic tag structure is:

Region	Attribute	Options	Weight ( $w_k$ )	Type
Top/Front	Length	6	2.25	Continuous
	Direction	8	2.0	Discrete
	Curly Level	4	1.0	Continuous
Side	Length	5	2.25	Continuous
	Curly Level	4	1.0	Continuous
Braid	Presence (Yes/No)	2	5.0	Discrete
	Count	4	2.0	Discrete
	Position	3	1.0	Discrete
	Type	5	1.0	Discrete

Table 1: Semantic Tag Design Specification with Weights

**Weight Rationale:** The weights  $w_k$  were determined through our iterative design process to reflect the relative importance of different attributes for human perception of hairstyle similarity. Braid presence receives the highest weight (5.0) as it represents a fundamental structural difference. Hair length attributes receive high weights (2.25) as they significantly impact visual appearance. Direction and braid count have moderate weights (2.0), while texture-related attributes (curliness) and fine-grained braid details receive lower weights (1.0).

The total theoretical tag space has  $6 \times 8 \times 4 \times 5 \times 4 \times 2 \times 4 \times 3 \times 5 = 460,800$  possible combinations, though only a subset of these correspond to actual avatar assets in our rendering system.

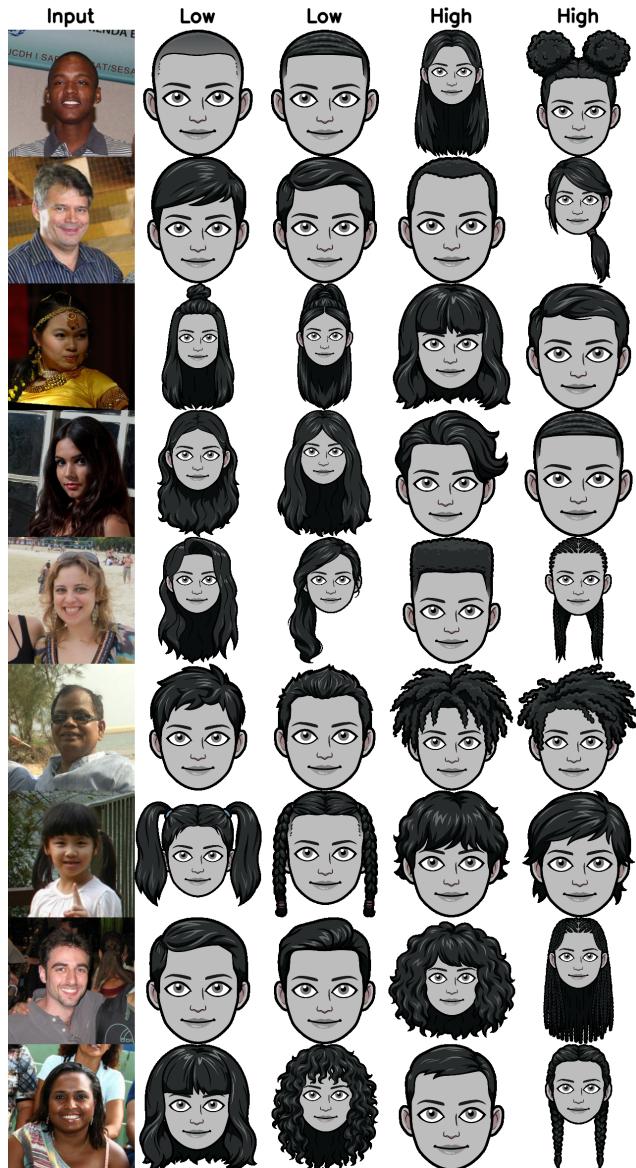
## A.5. Implementation Details

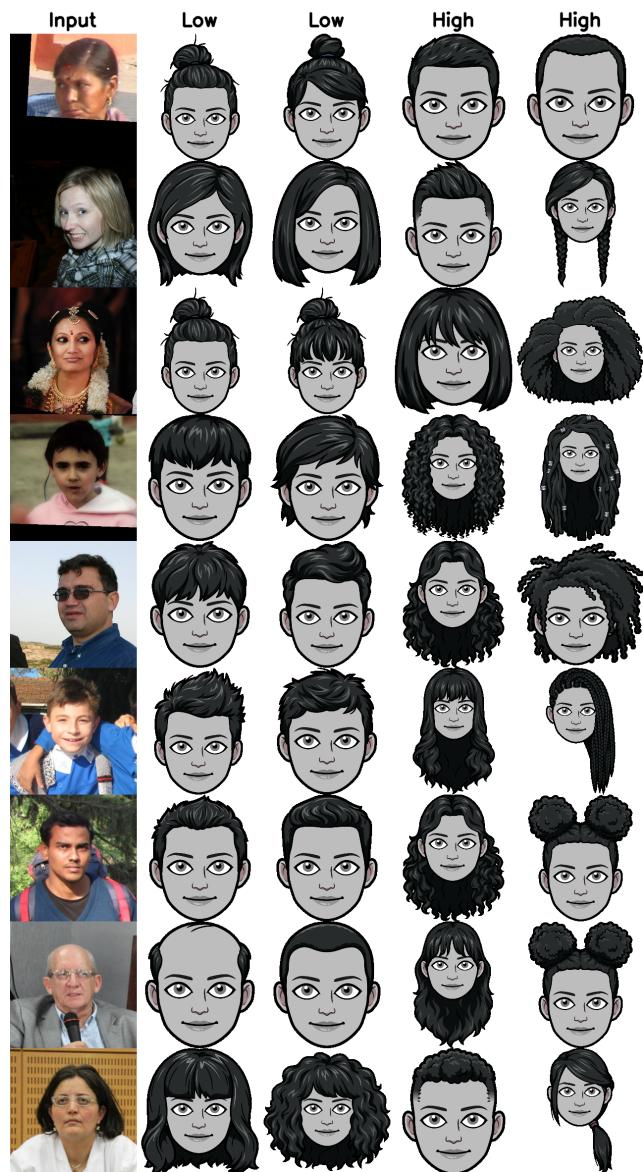
**Neural Architecture:** We use ResNet-50 as the backbone, pre-trained on ImageNet, with custom prediction heads for each attribute. Continuous attributes use linear output layers with sigmoid activation, while discrete attributes use linear layers with softmax activation.

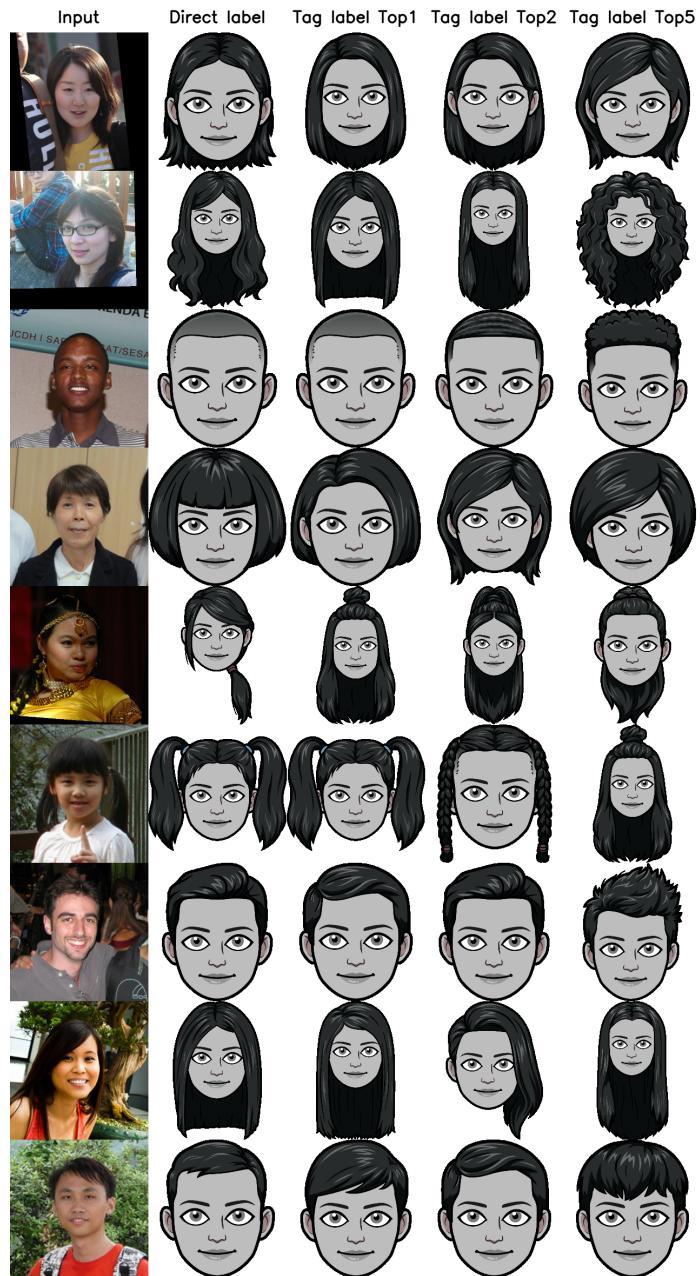
**Training Configuration:** We train for 100 epochs using Adam optimizer with learning rate  $\alpha = 0.001$ , batch size 32, and standard data augmentation (horizontal flip, rotation, color jitter).

**Symbolic Search Optimization:** The distance computation can be vectorized for efficiency. We precompute all avatar tag representations and use broadcasting operations to compute distances for all avatars simultaneously.

**Appendix B. Additional samples for Visualization of low and high distance samples (Fig3)**

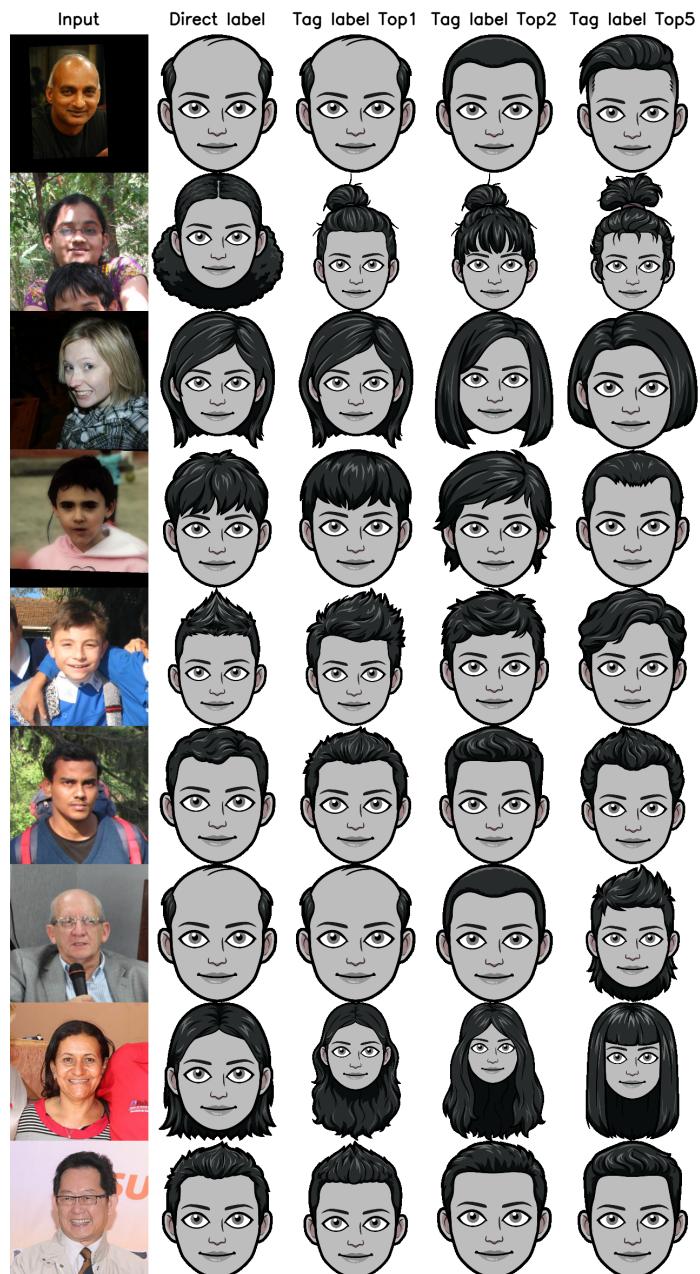


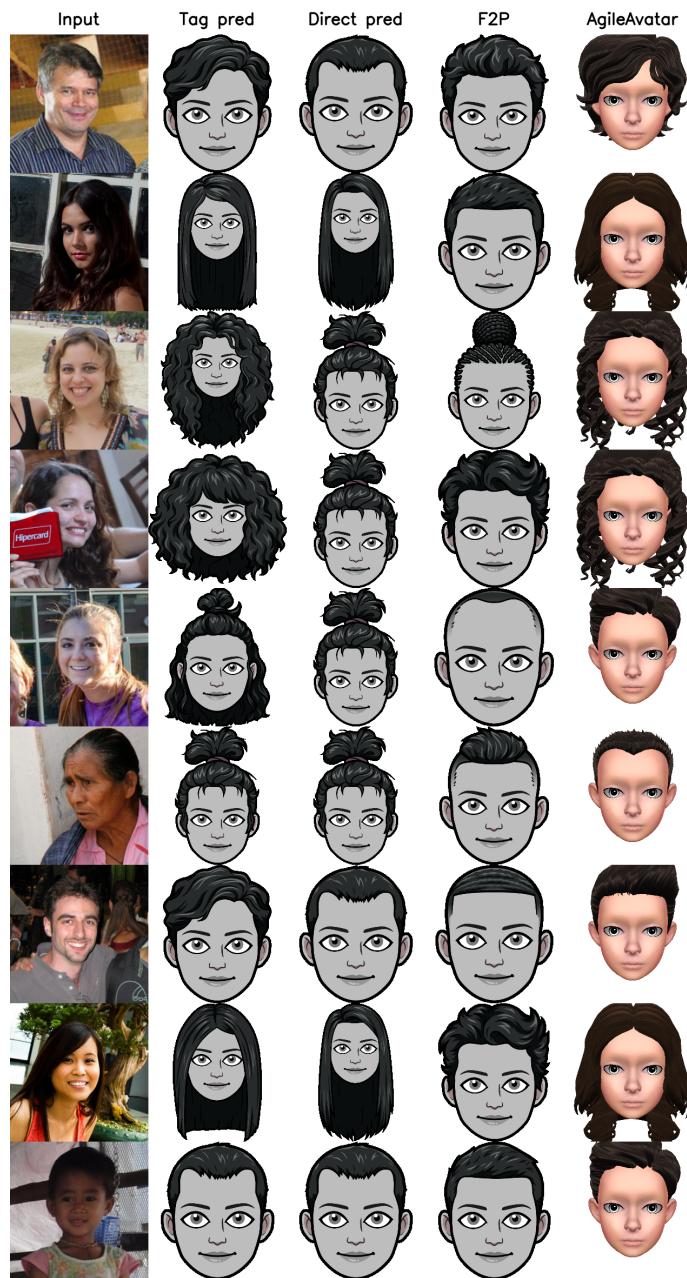


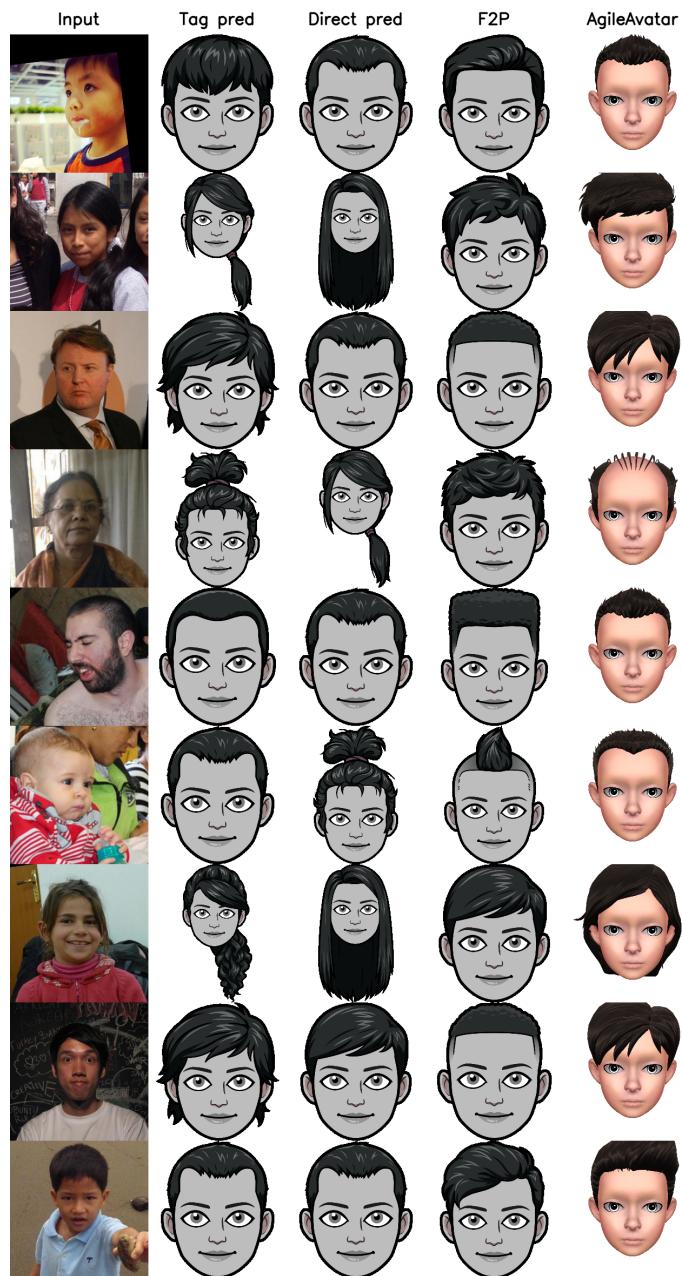
**Appendix C. Additional samples for Visual comparison of annotation results (Fig4)**

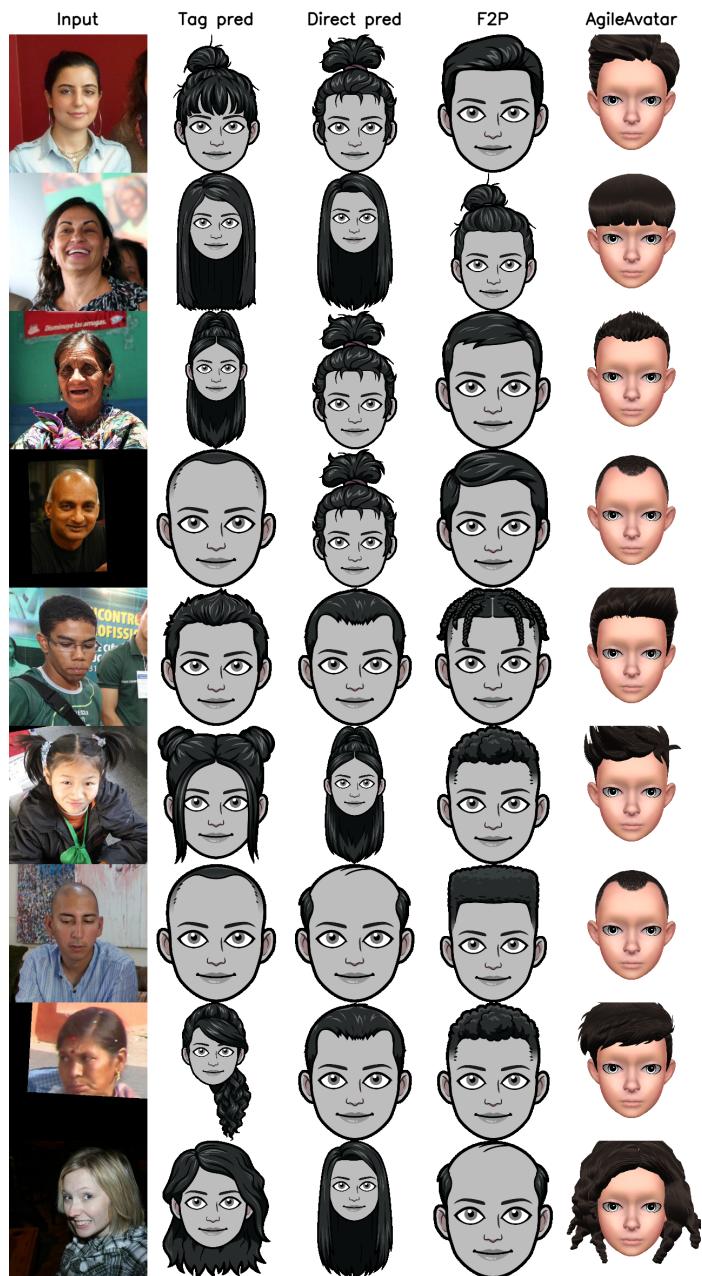






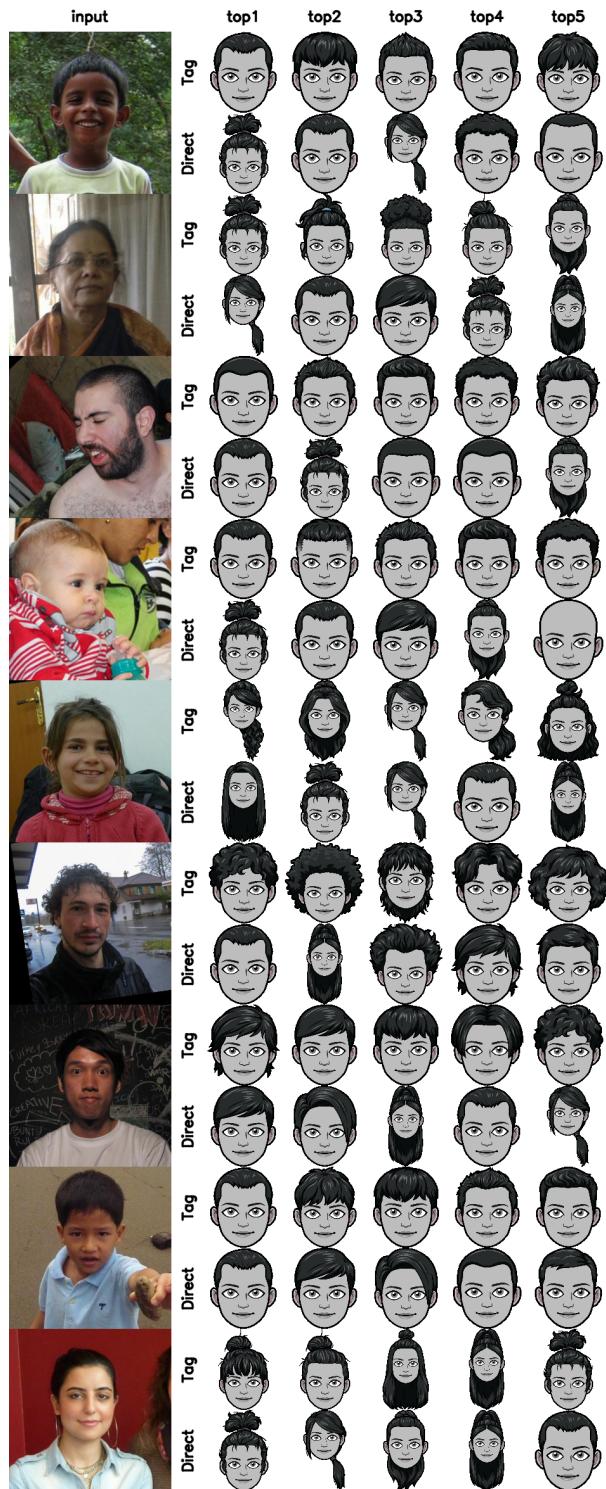
**Appendix D. Additional samples for Visual comparisons of model predictions (Fig5)**

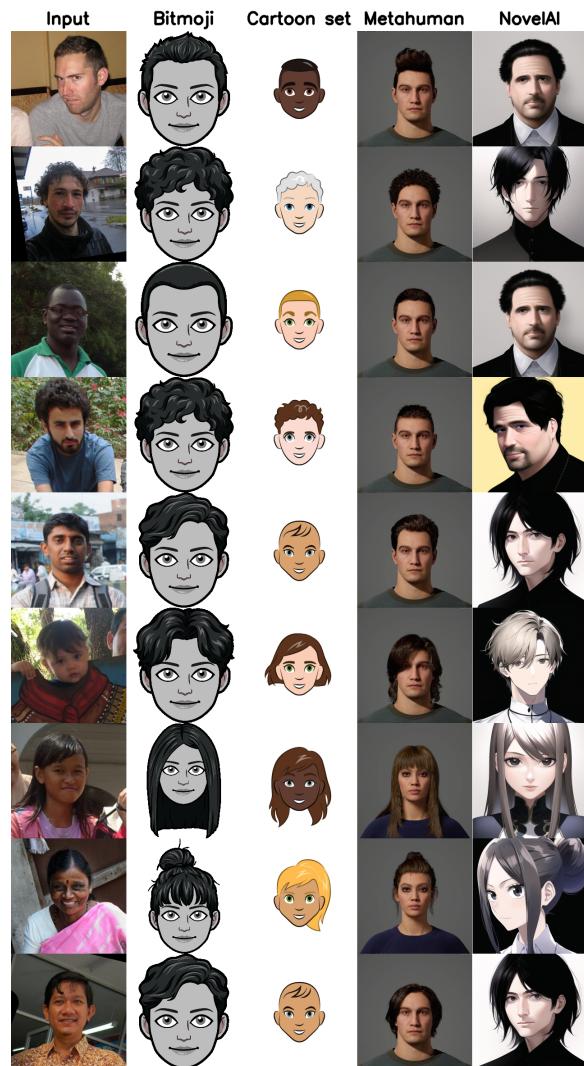


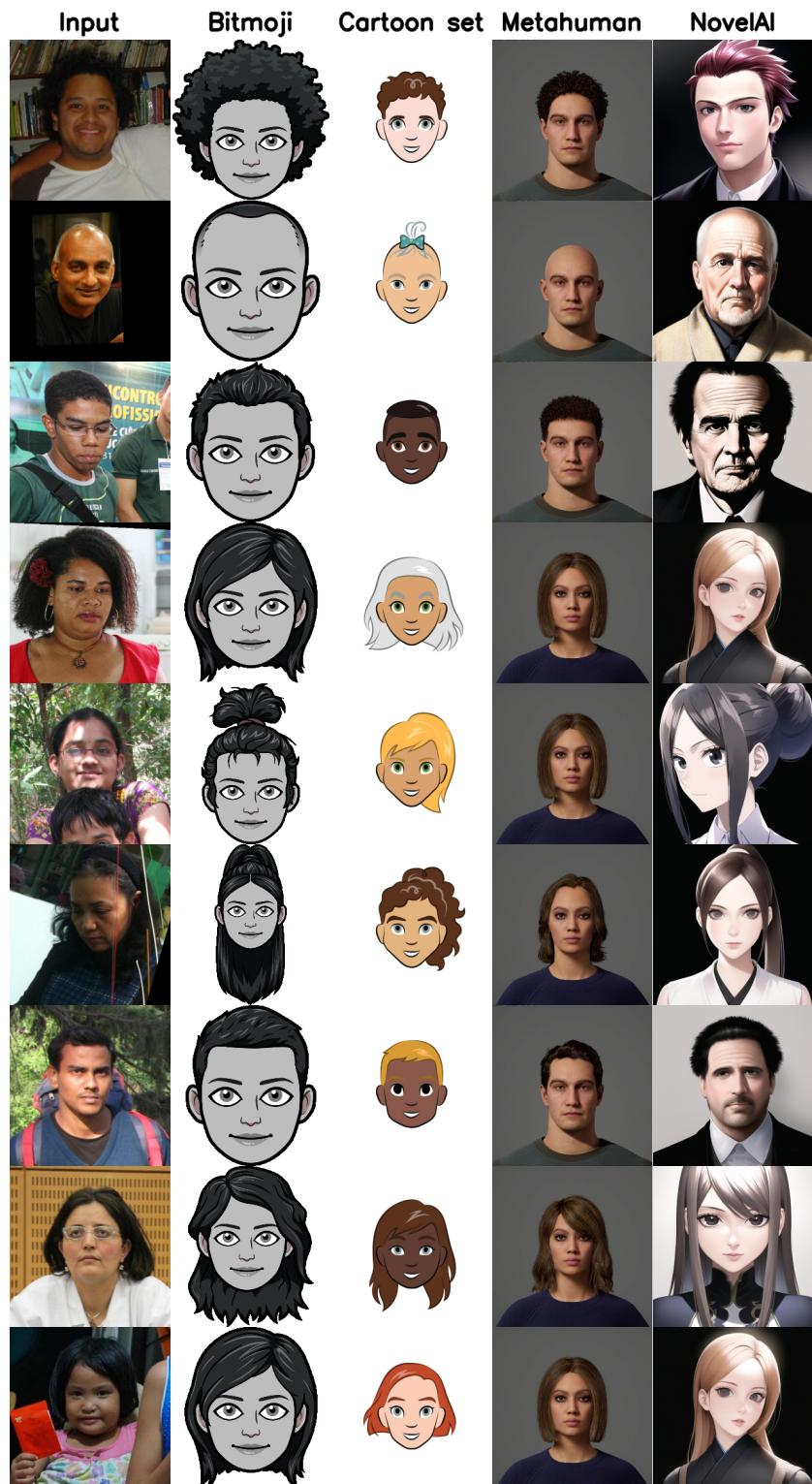


## Appendix E. Additional samples for Prediction consistency (Fig6)





**Appendix F. Additional samples for Easily expandable to other systems  
(Fig7)**



## Appendix G. Tag-based annotation page



**Top and front side:**

Length Direction Curl Level

**Side:**

Length Curl Level

**Braid:**

YesNo Braid Count Position

**Braid Style:**

Submit

Please annotate the length of top and front side hair. Focus on the absolute hair length, regardless of gender. (Ignore hair on the side of the head for now)

Please annotation hair on top of the head and bangs 

**no hair** **High hairline (Bald)** **Crew cut** **Short Hair** **Medium Hair** **Long Hair**

They have all the hair on top of their head, you can't see any hair.  
They have a high hairline, maybe suffers from hair loss.  
They have very short hair, like a bristly surface of a brush.  
Natural wind won't mess up this hair.  
Top side hair doesn't reach his eyebrows.  
Top side hair reaches eyebrow or longer.

Does this group satisfy your expect ?



Yess, Next



**Top and front side:**

Length Direction Curl Level

**Side:**

Length Curl Level

**Braid:**

YesNo Braid Count Position

**Braid Style:**

Submit

Please annotate the Direction of top and front side hair. (Ignore hair on the side of the head for now)

Please annotation hair on top of the head and bangs 

**Straight Downward** **Decline Downward** **Incline Upward** **Horizontal left or right** **Central parting** **Side parting** **Upward** **Backward**

Bang straight downward.  
Top side hair falls to one side at around 45 degrees.  
Top side hair sticks mostly upward at around 45 degrees.  
Hair comes from brushed mostly sideways.  
Hairstyle where the hair is separated at one side.  
Top side hair reaches eyebrow or longer.  
Hair standing upright or brushed so that it mostly sticks up.  
Hair pulled back, exposing forehead.

Does this group satisfy your expect ?



Yes, Next



**Top and front side:**

Length Direction Curl Level

**Side:**

Length Curl Level

**Braid:**

YesNo Braid Count Position

**Braid Style:**

Submit

Please annotate the Curl level of top and front side hair. (Ignore hair on the side of the head for now)

Please annotation hair on top of the head and bangs 

**Straight** **Wave** **Curly** **Coiled**

Naturally straight hair.  
Not quite straight and not completely wavy. Also, hair of someone who has wavy hair, or well styled.  
Ranges from a light curl to tight, curly waves.  
Fine and thin or wavy and coarse, tightly packed coils.

Does this group satisfy your expect ?



Yes, Next

# LIU CHENG SANG LIU DAVIS



**Top and front side:**

**Side:**

**Braid:**



Please annotate the length on side of the head. Focus on the absolute hair length, regardless of gender. (ignore hair on the top of the head for now)

**Please annotation hair on side of the head** 

Very Short	Short	Medium short hair	Medium Long hair	Long hair
Hair mostly shaved on the side. Briefly, like the surface of a book. For example, undercut.	Natural and can't mess up this hair.	Longer than short hair, but without covering the ear.	Length ranges from covering the ear to reaching the shoulder.	Hair can stay on the shoulders and longer.

**Does this group satisfy your expect ?**





**Top and front side:**

**Side:**

**Braid:**



Please annotate the Curl level of top and front side hair. (ignore hair on the side of the head for now)

**Please annotation hair on side of the head** 

Straight	Wave	Curly	Coiled
Naturally straight hair.	Not quite straight hair. It's slightly wavy or curly. Also includes hair that is messy hair or wet messy hair.	Ranges from a light curl to a tightly coiled tendril. Looks like a beach ball.	Fine and thin or coarse and chunky, with densely packed coils.

**Does this group satisfy your expect ?**





**Top and front side:**

**Side:**

**Braid:**



The person used some threads, hair band to tie up his hair.

**Please annotation their braid, click no braid if they don't have any** 

No	Yes
They are not wearing a braid.	They are wearing a braid.

**Does this group satisfy your expect ?**





**Top and front side:**

Length:  Direction:  Curl Level:

**Side:**

Length:  Curl Level:

**Braid:**

Yes/No:  Count:  Position:



How many braids does the person have?

Please annotation their braid, click no braid if they don't have any



No braid  
They have No braid.



Single  
There is only one braid. For example ponytail.



Double  
There are two braids. For example double buns.



Multiple  
There are more than two braids. For example Dreadlock.

Does this group satisfy your expect ?





**Top and front side:**

Length:  Direction:  Curl Level:

**Side:**

Length:  Curl Level:

**Braid:**

Yes/No:  Count:  Position:



The braid position is high or low.

Please annotation their braid, click no braid if they don't have any



No braid  
They have No braid.



Low  
They have the braid on the lower side of his head.



High  
They have the braid on the upper side of his head.

Does this group satisfy your expect ?





**Top and front side:**

Length:  Direction:  Curl Level:

**Side:**

Length:  Curl Level:

**Braid:**

Yes/No:  Count:  Position:



What kind of braid is worn?

Please annotation their braid, click no braid if they don't have any



No braid  
They have No braid.



Ponytail  
Hair tied at the back of the head causing it to hang down like a tail of pony.



Plaited  
One or more braids, parallel or interlaced.



Deadstock  
A narrow rope-like braid formed by matting, braiding, or twisting.



Bun  
Hair wrapped in a circular coil around the head.

Does this group satisfy your expect ?



LIU CHENG SANG LIU DAVIS