

CosmicMan: A Text-to-Image Foundation Model for Humans

Supplementary Material

Shikai Li*, Jianglin Fu*, Kaiyuan Liu*, Wentao Wang* Kwan-Yee Lin†, Wayne Wu†,
Shanghai AI Laboratory

{lishikai, fujianglin, wangwentao}@pjlab.org.cn

1154864382@mail.dlut.edu.cn linjunyi9335@gmail.com wuwenyan0503@gmail.com

Abstract

This supplement serves to further enrich the discourse established in our main paper. In Sec. 1, we detail the release of the CosmicMan-HQ 1.0 dataset and foundation models. In Sec. 2, we present additional elaboration on the scalable data production pipeline – Annotate Anyone. Then, in Sec. 3, we delve into the proposed training framework (i.e., Daring) with additional evaluation and illustrations. Lastly, we present a broader range of qualitative results, including comparisons with state-of-the-art methods, more examples of generated samples, extensive ablation studies, and potential application demonstrations in Sec. 4.

1. Release

We seriously treat the license and privacy issues and follow a rigorous legal review in our institute. CosmicMan-HQ 1.0 with all of the annotations will be released step by step. People in the dataset are anonymized without additional private or sensitive metadata. All released data are free for research use only. We will release all the codes and weights of the proposed human-specialized foundation model trained on CosmicMan-HQ 1.0.

2. Annotate Anyone

As introduced in the main paper, the whole data production paradigm comprises two parts that benefit from the cooperation of human effort and AI models – a flowing data sourcing for data collection, and human-in-the-loop data annotation. In this section, we unfold the details of Annotate Anyone. We first complement the data filtering step in Sec. 2.1. Then, we provide the pseudo-code for fine-tuning the VLM model used in Annotate Anyone, accompanied by an exhaustive explanation (Sec. 2.2). Besides, we conduct experiments to demonstrate the effectiveness of the proposed human-in-the-loop annotation pipeline in Sec. 2.3. More-

Algorithm 1 Annotate Anyone:

Iteratively improve VLM model to annotate image

Input: I ▷ Images in Data Pool
 $I_0 \leftarrow$ sample test set from I ▷ K images
 $A_0 \leftarrow$ label I_0 by VQA pretrained model (IB_0)
 $M_0 \leftarrow$ label I_0 by human
 $Acc \leftarrow$ Eval(A_0, M_0) ▷ Acc for all label
 $i = 0$
while $Acc < 85\%$ **do**
 $i \leftarrow i + 1$
 $I_i \leftarrow$ sample K images from data pool
 $\tilde{M}_i \leftarrow$ select labels from M_i with $Acc < 85\%$
 $IB_{i+1} \leftarrow$ finetune IB_i with $(\cup_1^i \tilde{M}_i, I_i)$
 $A_{i+1} \leftarrow$ label I_0 by (IB_{i+1})
 $Acc \leftarrow$ Eval(A_{i+1}, M_0)
end while
 $(I, A) \leftarrow$ update all image label in data pool using IB_{i+1}

over, we provide a detailed exposition of all 70 questions we used on VLM for obtaining the description of an image, as presented in Sec. 2.4. We also list the corresponding attribute labels extracted from the output of VLM models. Finally, a detailed statistical analysis (Sec. 2.5) and additional data samples from our dataset CosmicMan-HQ 1.0 (Sec. 2.6) are presented.

2.1. Data Filtering

As mentioned in the main text, we use three academic datasets, LAION-5B [22], SHHQ [5], and DeepFashion [16] as part of data origin. When distilling LAION-5B dataset, we extensively utilize the meta information provided by LAION-5B itself to perform an initial screening of the data. We filter out a significant amount of images that contain the “Not safe for work (nsfw)” label or have a watermark score higher than 0.5. The data source of SHHQ and DeepFashion is relatively concentrated, consisting exclusively of watermark-free fashion / studio-shot data. Thus,

we omit this step for those two datasets. Before fetching images from the Internet, we excluded the data origins of LAION-5B, SHHQ, and DeepFashion. Our aim is to minimize crawling data from the same source as much as possible. We also remove duplicated images by encoding images into fixed-length hash values using perceptual hashing algorithms [9]. Images with hamming distances smaller than 3 are considered duplicates, and one of them will be removed to avoid redundancy.

To further remove fake-people images (*e.g.*, cartoon characters, mannequin models, and generated images), we fine-tuned Eva-CLIP [23] with Human-Art [12] and sets of fake images. The fine-tuned model with 91% accuracy is used to detect images containing fake people. Next, image quality assessment metrics (LIQE [26], IFQA [11], and HPSv2 [25]) are applied to evaluate image quality at both face and global levels, aiding in refining the data pool further. Images with subpar overall quality and face quality are eliminated. Further, we utilize YOLOv7 [24] to filter out images without humans or those containing more than one individual. Images where the detected largest face has a resolution smaller than 224×224 or where the entire image measures less than 640×1280 are discarded as well.

2.2. Human-in-the-loop Data Annotation

This section presents the pseudo-code of the annotation algorithm (see Algorithm. 1) mentioned in Sec. 3.2.2 of the main text for Annotate Anyone. A detailed step-by-step explanation will be provided below.

The overall objective of this process is to achieve collaboration between AI and human to provide descriptions for all images (denoted as I) within the entire data pool. We start the workflow by randomly selecting a fixed test set I_0 , consisting of K images from the data pool I . I_0 will be asked 70 meticulously crafted vision questions (see Tab. 1 for reference) by the pretrained model IB_0 [3] and obtain corresponding answers A_0 . Note that, inspired by the annotation approach used in Fashion datasets [10, 16], we transform the uncontrollable, attributes highly entangled captioning tasks into detailed question-answering tasks to get A_0 . This approach results in structured label answers, proceeding from coarse to finer granularity, that can ease evaluation and provide more accurate and fine-grained region-specific labels. Then, we ask a professional annotation team to manually label the images in I_0 based on these 70 questions, to form the attribute dictionary M_0 as the ground-truth of the test set. For every single question, we evaluate the accuracy of the pretrained model IB_0 by comparing the label in A_0 and M_0 .

The label accuracy for every attribute corresponding to every question will be documented and utilized to assess whether the current (pretrained and fine-tuned) model’s judgment for that attribute meets the criteria. If there is

one label’s accuracy less than 85%, the fine-tuning loop of AI model IB starts. Concretely, We will randomly sample another K images from the data pool, and ask the annotation team to label the attributes whose model accuracy is less than 85%. After each round of manual annotation, the newly manual-labeled data is combined with the previously annotated data and utilized for fine-tuning the IB_i model. During training, we use “<Image >: {}. Question: {} Answer:” as a template to construct fine-tuning dataset. The fine-tuned model IB_{i+1} will be utilized to perform inference on the test set I_0 for evaluation to get Acc . This fine-tuning process will repeat until the overall label accuracy achieves 85%. The final model IB_{i+1} is used to annotate all the images in the data pool, and results image-label pairs (I, A) .

2.3. Effectiveness of Annotate Anyone

In this section, we conduct experiments to assess the effectiveness of reducing manual annotation efforts in our Annotate Anyone pipeline.

We run four sets of model fine-tuning with a step-wise augmentation of manually labeled data. In every iteration, we randomly sampled 1000 images (K in Algorithm. 1 is set to 1000) from I in the data pool. Leveraging evaluation metrics (Acc) from the previous model, we progressively decrease the need for manual labeling by excluding already qualified attributes (where the Acc is higher than 85%). The annotated data accumulates and is utilized for fine-tuning the subsequent model.

As shown in bar charts of the first iteration in Fig. 2, due to the random sampling from the data pool, the inherent attribute distribution of sampled images naturally exhibits a long-tail pattern similar to that of CosmicMan-HQ 1.0 (Fig. 3). The accuracy of different attributes also correlates to the number of training data. Then in the following iterations, as we only augment the attributes that did not meet the accuracy threshold, the long-tail distribution of the training data is alleviated, as shown in the bar charts of Fig. 2. Specifically, with the progression of iterations, the number of labels requiring manual annotation on each image gradually decreases. Results in the figure showcase that the tail-end labels have improved over iterations while the existing qualified labels maintain their accuracy. The overall accuracy of the final fine-tuned model has significantly increased from 56.0% in the pre-training phase to 85.3%. We have temporarily halted this flywheel because the average accuracy has reached 85%, but it can continue run until the accuracy of every single attribute meets the standard.

2.4. Label Protocol

In Tab. 1, we present a series of questions related to each part from the human parsing map, resulting in a total of 70 question-answer pairs to describe a human image.

Table 1. **The Detailed Questions Asked to Describe Every Single Image.** The index 0-2 in Acc represents the corresponding questions is evaluated as **Acc_{obj}**, **Acc_{tex}**, and **Acc_{shape}**.

	Q _{idx}	Acc	Questions	Attributes
Overall	1	0	what is the gender of the person in the image?	male, female
	2	0	what is the country of the person in the image?	Indian, Latino, Middle Eastern, ...
	3	0	what is the age of the person in the image?	teenager, adult, elderly, ...
	4	0	what is the body shape of the person in the image?	fit, skinny, obese, muscular
	5	0	what is the background of the person in the image?	
	6	0	what is the overall-style of the person in the image?	fashion, documentary, portrait, others
Hair	7	0	is the hair of the person visible in the image?	yes, no
	8	-	what is the hair color of the person?	brown, gray, violet, ...
	9	1	what is the hairstyle of the person?	wavy, ponytail, straight, ...
	10	2	what is the hair length of the person?	below chest, bald, bob, ...
Top Clothings	11	0	does the person wears any tops?	yes, no
	12	0	what kind of top does the person wear?	vest, blouse, hoodie, ...
	13	1	what is the pattern of the tops?	graphic, printed, stripes, ...
	14	1	what is the material of the tops?	linen, fur, lace, ...
	15	2	what is the sleeve length of the tops?	short, medium, sleeveless, ...
	16	2	what is the top length of the tops?	crop, normal, tunic
	17	2	what is the collar shape of the tops?	square, v-shape, collar, ...
	18	-	are the top clothing {random color}?	yes, no
Bottom Clothings	19	0	does the person wears any bottoms?	yes, no
	20	0	what kind of bottom does the person wear?	shorts, sweatpants, skirt, ...
	21	1	what is the pattern of the bottoms?	graphic, printed, stripes, ...
	22	1	what is the material of the bottoms?	linen, fur, lace, ...
	23	2	what is the length of the bottoms?	short, medium, sleeveless, ...
	24	2	what is the bottom shape of the bottoms?	straight, tapered, wide-leg, ...
One-piece Outfits	25	-	are the bottom clothing {random color}?	yes, no
	26	0	does the person wears any one-piece outfits?	yes, no
	27	0	what kind of one-piece outfit does the person wear?	bathrobe, jumpsuit, dress, ...
	28	1	what is the pattern of the one-piece outfits?	graphic, printed, stripes, ...
	29	1	what is the material of the one-piece outfits?	linen, fur, lace, ...
	30	2	what is the sleeve length of the one-piece outfits?	short, medium, sleeveless, ...
	31	2	what is the collar shape of the one-piece outfits?	square, v-shape, collar, ...
Coats	32	2	what is the length of the one-piece outfits?	short, medium, sleeveless, ...
	33	2	what is the shoulder exposure level of the one-piece outfits?	one-shoulder, off-shoulder
	34	0	does the person wears any coats?	yes, no
	35	0	what kind of coat does the person wear?	cape, trench coat, blazer, ...
	36	1	what is the pattern of the coats?	graphic, printed, stripes, ...
	37	1	what is the material of the coats?	linen, fur, lace, ...
	38	2	what is the coat length of the coats?	short, medium, maxi
	39	2	what is the collar shape of the coats?	square, v-shape, collar, ...
Special Clothings	40	-	are the coat {random color}?	yes, no
	41	0	does the person wears any special clothings?	yes, no
	42	0	what kind of special clothing does the person wear?	Hanfu, Saree, cosplay, ...
Shoes	43	2	what is the sleeve length of the special clothing?	short, medium, sleeveless, ...
	44	0	does the person wears any shoes?	yes, no
	45	0	what kind of shoe does the person wear?	sneakers, flip flops, loafers, ...
	46	1	what is the pattern of the shoes?	graphic, printed, stripes, ...
	47	1	what is the material of the shoes?	linen, fur, lace, ...
Bags	48	2	what is the boots length of the shoes?	ankle, mid-calf, knee-high, ...
	49	-	are the shoes {random color}?	yes, no
	50	0	does the person wears any bags?	yes, no
Hats	51	0	what is the type of the bags?	tote bag, handbag, wallet, ...
	52	1	what is the material of the bags?	linen, fur, lace, ...
	53	0	does the person wears any hats?	yes, no
Socks	54	0	what is the type of the hats?	beret, sun hat, helmet, ...,
	55	1	what is the material of the hats?	linen, fur, lace, ...
Other Accessories	56	1	what is the pattern of the socks?	graphic, printed, stripes, ...
	57	1	what is the material of the socks?	linen, fur, lace, ...
	58	1	what is the pattern of the belts?	graphic, printed, stripes, ...
	59	1	what is the material of the scarf?	linen, fur, lace, ...
	60	1	what is the pattern of the scarf?	graphic, printed, stripes, ...
Headwear	61	1	what is the material of the ties?	linen, fur, lace, ...
	62	1	what is the pattern of the ties?	graphic, printed, stripes, ...
	63	0	does the person wears any headwear?	yes, no
	64	0	what kind of headwear does the person wear?	headband, headscarf, veil
	65	1	what is the pattern of the headband?	graphic, printed, stripes, ...
	66	1	what is the material of the headband?	linen, fur, lace, ...
	67	1	what is the pattern of the headscarf?	graphic, printed, stripes, ...
	68	1	what is the material of the headscarf?	linen, fur, lace, ...
	69	1	what is the pattern of the veil?	graphic, printed, stripes, ...
	70	1	what is the material of the veil?	linen, fur, lace, ...





				
Raw Text	Flora and Fauna Dress	Best of Christmas Project	Alexa Mansour as Hope - The Walking Dead: World Beyond _ Season 1, Episode 1 - Photo Credit: Zach Dilgard/AMC	-
BLIP	a woman wearing a leopard print dress and a hat	there is a little girl sitting on a rug with a toothbrush	a woman sitting in a chair in a kitchen	a woman in a short dress holding an umbrella
IntructBlip	the leopard off the shoulder dress	The person in the picture is a young girl sitting on the floor next to a Christmas tree, playing with a toy.	The image features a woman sitting on a chair in a living room. She is wearing jeans and a sweater, with her hands resting on her knees. There are several objects scattered around the room, including a vase, a cup, and a bottle of wine.	The image depicts a young woman standing under an umbrella in a park. She is wearing a black skirt and white top, while holding an umbrella to protect herself from the sun.
Ours	A nearly full-body shot, a Caucasian adult female, fit, indoor, wavy above chest brown hair, woven hat, cotton graphic medium sleeve short off-shoulder dress	A nearly full-body shot, a Caucasian child female, fit, indoor, cotton short sleeve dress in red, long sleeve cotton stripes shirt, bob blonde above shoulders hair	A nearly full-body shot, an middle eastern female, fit, indoor, black wavy hair above the chest. Cotton short sleeve t-shirt in gray with a v-shape neckline, blue denim jeans with maxi length	A full-body shot, an Asian adult female, outdoor, black straight above chest hair, a black silk shirt with long sleeve and collar, cotton white and short pleated skirt, black leather loafers, white cotton socks

Figure 1. **Data and Annotation Samplings.** Sample images with captions obtained from different methods: the original text uploaded along with the images, captions generated by BLIP [13], and InstructBLIP [3] pretrained models, and our attribute-based text. Here we use different colors to highlight the unrelated, wrong, and coarse and vague descriptions.

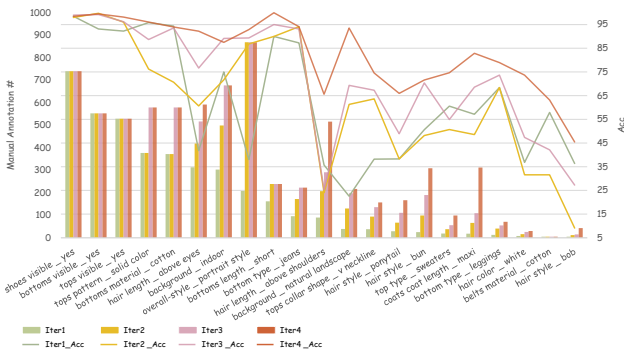


Figure 2. **Effectiveness of Annotate Anyone.** The bar charts represent the manual annotation counts for different attributes in 1000 images for each iteration, and the solid lines depict the accuracy of each attribute under the current model. Better zoom in for details.

2.5. Statistics of CosmicMan-HQ 1.0

To visually demonstrate the superiority of our dataset in terms of caption granularity and accuracy, Fig. 1 presents images from CosmicMan-HQ 1.0 paired with annotations from different sources. As shown, the text, including the raw caption from original websites and generated caption

from the BLIP [13], often lacks descriptive details. Raw text is often unrelated to the image content, while text generated by BLIP is correct in general but vague, and the captions from the InstructBLIP [3] have lower accuracy in providing detailed captions, particularly in attribute-level descriptions. For example, InstructBLIP incorrectly stated that the woman is wearing a sweater instead of a short-sleeved shirt, as shown in the description of the third image in Fig. 1. Compared to these captions, our method provides comprehensive and accurate descriptions by transforming fine-grained labels using the fine-tuned IB model.

Following that, We delve into an in-depth exploration of the data distribution within the dataset. The statistical analysis of the overall image and human-centric attributes across CosmicMan-HQ 1.0 is displayed in the left half of Fig. 3. The right half offers examples of fine-grained attributes related to “top clothing”, which further demonstrates the granularity of our annotation. The histogram presents a wide coverage of various attributes and a natural long-tail distribution, which reflects our dataset’s consistency with the data in the real world.

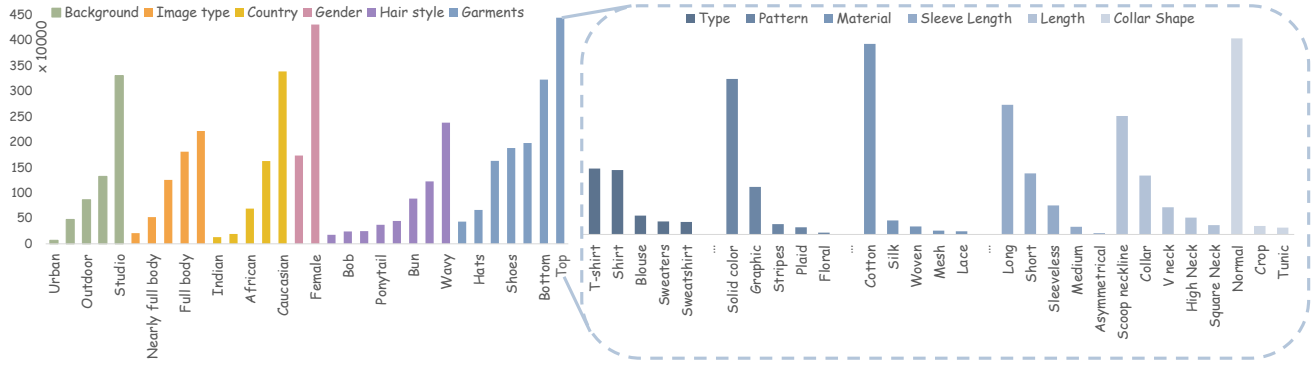


Figure 3. **CosmicMan-HQ 1.0 Statistics.** The left part shows the global attributes and garment types of the person in the image, and the right section displays attributes examples of top garments.



Figure 4. **CosmicMan-HQ 1.0 Image Examples.** The examples display the dataset’s wide-ranging diversity, encompassing different aspects like age, ethnicity, and clothing styles.

2.6. More Dataset Samples

In Fig. 4, more image examples are randomly sampled from CosmicMan-HQ 1.0, which showcases the dataset’s inherent diversity across various dimensions, such as age, ethnicity, and garments.

3. Daring

In this section, we primarily focus on the additional analysis of our proposed training framework, *Daring*. Regarding the Data Discretization, we supplement with an ablation

study on the type of captions, validating the effectiveness of the decomposed text space in Sec. 3.1. We then provide a comparison on the two terms of the proposed HOLA loss in Sec. 3.2. By presenting the cross-attention maps of the proposed model, we elucidate that our method improves the text-image alignment for dense concepts in Sec. 3.3. We also conduct an experimental evaluation of optimization strategies involving different loss functions in Sec. 3.4.

3.1. Evaluation of Data Discretization

Our framework first decomposes both the text and image into several groups of pair data. The underlying motivation is that we hypothesize the descriptions are decomposable and finite for describing human images, since they are inherently associated with human body structure, and each concept can correspond to a particular body region. To evaluate the effectiveness of Data Discretization, we conduct an experiment on decomposed and continuous text space as shown in Tab. 2. Of note, to cost-effectively get natural continuous text, we use GPT-4 to generate 5-10 templates for each set of labels in the human structure, such as: “{type} with {pattern} made of {material}, {sleeve_length} sleeves, {collar_shape} neckline, {length} length, {shoulder_exposure_level} exposure” for “One-piece Outfits” group. The natural descriptions of each group are concatenated to get a sentence for training and testing. Since the text encoder of SD can only handle 77 tokens, we construct another test set with short captions. For each sample within the original test set, when the number of tokens within the natural description is greater than 77, we randomly dropout descriptions of certain body parts. It can be seen that with the decomposed text data, the \mathbf{Acc}_{all} exhibits a significant improvement. This is because decomposed text space ensures that all important details are accurately and clearly included and does not introduce ambiguity or unnecessary context. While data discretization gives an additional boost in performance, the best performance is achieved when \mathcal{L}_{HOLA} and data discretization are combined. This illustrates that as long as keys K in attention blocks are initially decomposable and semantically guided, it may not be necessary to optimize latent code using cross-attention maps in the inference or add complex modules to SD’s original architecture for dense concepts alignment.

Table 2. **Ablation on \mathcal{L}_{HOLA} and Data Discretization.** The best results are marked with **Red**.

\mathcal{L}_{HOLA}	Discretization	$\mathbf{Acc}_{obj}\uparrow$	$\mathbf{Acc}_{tex}\uparrow$	$\mathbf{Acc}_{shape}\uparrow$	$\mathbf{Acc}_{all}\uparrow$
		74.6	88.4	56.2	73.1
✓		76.8	89.3	59.3	75.1
	✓	82.4	91.3	66.5	80.1
✓	✓	85.5	92.2	71.4	83.1

3.2. Evaluation of HOLA Loss

The proposed loss \mathcal{L}_{HOLA} for guiding the model to align with the region-level captions is composed of two terms. The first term introduces the overall spatial structure and single concept, while the second term reduces the ambiguities of outfit-level descriptions. For example, patterns exhibited on clothing typically manifest in specific areas rather than uniformly across the entire garment. Conse-

quently, these patterns do not necessitate alignment across the entire clothing piece. Here we also provide quantitative results in Tab. 3. Baseline means the model that only fine-tuned on CosmicMan-HQ-1M. The employment of “term1” markedly enhances all semantic accuracy. Subsequent incorporation of “term2” yielded a more pronounced enhancement in \mathbf{Acc}_{tex} due to the fact that certain outfit-level texture attributes such as patterns don’t need to be aligned with the whole semantic region.

Table 3. **Ablation on Loss Terms in \mathcal{L}_{HOLA} .** Term1 pushes the high response region of each single concept feature. Term2 helps reduce the ambiguities of outfit-level descriptions.

Methods	$\mathbf{Acc}_{obj}\uparrow$	$\mathbf{Acc}_{tex}\uparrow$	$\mathbf{Acc}_{shape}\uparrow$	$\mathbf{Acc}_{all}\uparrow$
Baseline	87.3	77.4	59.3	74.6
+ term1	91.3	84.7	65.6	80.5
+ term2	91.7	85.7	66.1	81.2

3.3. Evaluation of Cross-Attention Maps

We also provide the comparison of the cross-attention maps across various models, as illustrated in Fig. 5. In summary, our evaluation indicates that CosmicMan-SDXL exhibits enhanced precision in the generation of graphic t-shirt and black sneakers. This is further corroborated by our detailed examination of the attention maps for “Top Clothing”. Notably, the attention map corresponding to the “graphic” attribute demonstrates a more accurate activation in the top region, particularly when compared with the attention map extracted from SDXL. Furthermore, a comprehensive comparison of all attributes shows that CosmicMan-SDXL not only achieves higher semantic accuracy, but also achieves more accurate semantic activation in the corresponding attention map than the model fine-tuned on CosmicMan-HQ.

3.4. Evaluation of Optimization Strategies

In addition to appending a loss term through linearization and adjusting scalar coefficients, we conduct additional experiments employing a multi-objective optimization technique, Random Loss Weighting (RLW) [14], on CosmicMan-SD. The results presented in Tab. 4 reveal that both using RLW directly and combining it with a weighted loss term lead to a decline in model performance.

Table 4. **Ablation on Optimization Strategies.** The best results are marked with **Red**.

RLW	β	FID↓	$\mathbf{Acc}_{obj}\uparrow$	$\mathbf{Acc}_{tex}\uparrow$	$\mathbf{Acc}_{shape}\uparrow$	$\mathbf{Acc}_{all}\uparrow$
✓		55.5	83.7	76.7	60.8	73.6
✓	✓	39.9	89.9	81.7	63.2	78.2
	✓	36.8	91.7	85.7	66.1	81.2

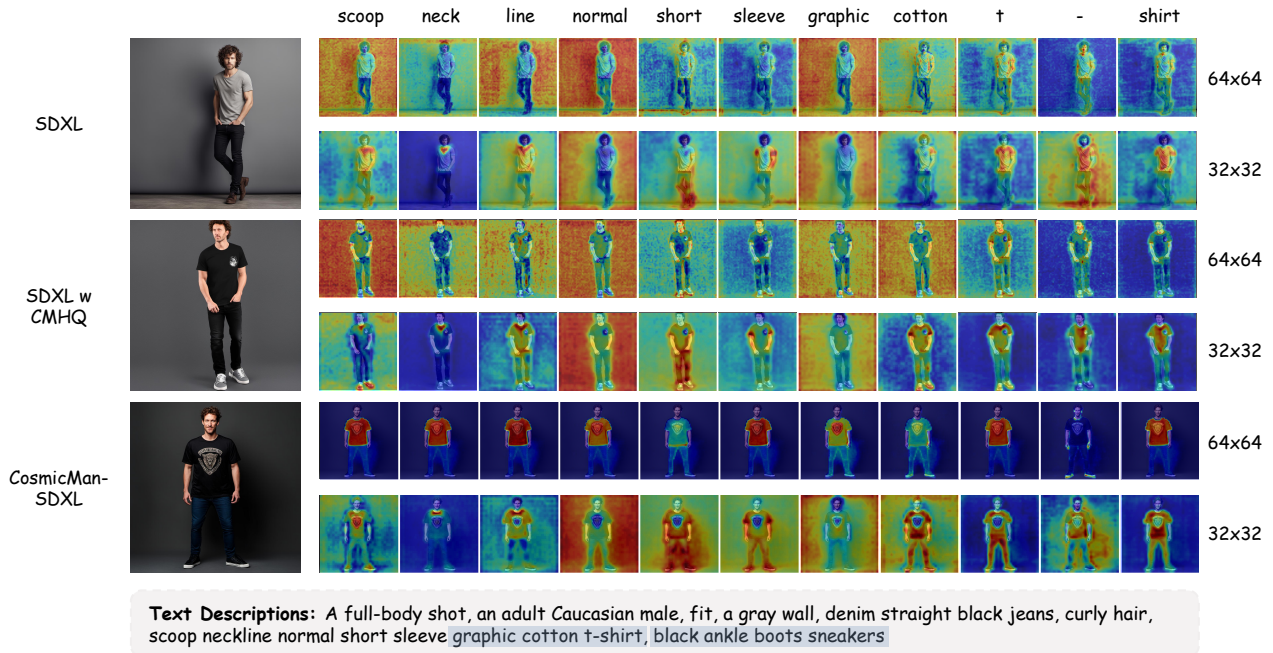


Figure 5. **Visualization of Cross-attention Maps.** We show cross-attention maps in SDXL-base for two different resolutions (map of 64×64 shape with $16 \times$ down-sampling and map of 32×32 shape with $32 \times$ down-sampling), in a similar way to the Prompt-to-Prompt [6] visualization method. For a clearer comparison, we further resize them to the original image size.

4. Experiments

To more vividly demonstrate the performance of our text-to-image foundation model, CosmicMan, this section provides an extensive array of qualitative results. In Sec. 4.1 and Sec. 4.2, we begin by detailing our experimental procedures and evaluation metric. We then present the comparative qualitative results against the state-of-the-art models in Sec. 4.3, followed by more visual results of the ablation studies in Sec. 4.4. In Sec. 4.5 and Sec. 4.6, we provide further analysis on the visual results with different granularity, as well as different concepts, respectively. Sec. 4.7 provides a fairness comparison on an unseen dataset. In Sec. 4.8, we show the superiority of our foundation model over Stable Diffusion pretrained model on two downstream applications. Finally, we provide additional visualizations of our model in Sec. 4.9.

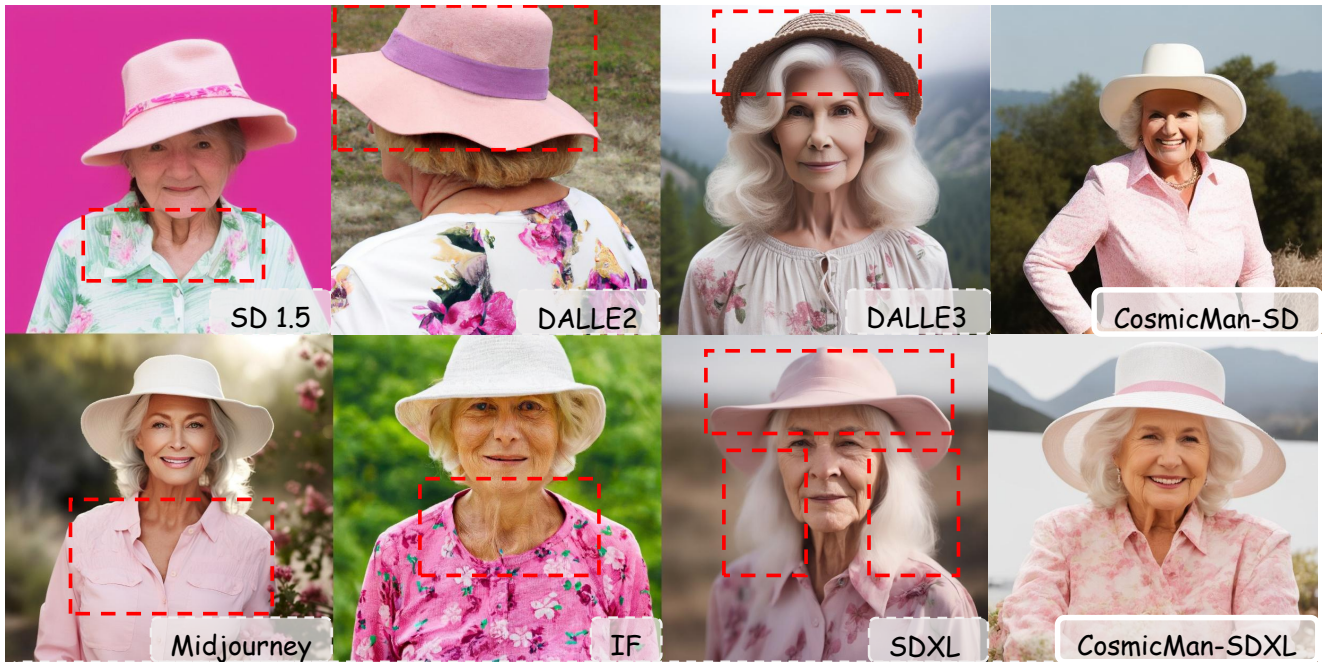
4.1. Experiment Setting

Our foundation models CosmicMan contains two versions: CosmicMan-SD based on SD-1.5 [21], and CosmicMan-SDXL based on SDXL [19]. The DDPM Scheduler [7] and EulerAncestralDiscreteScheduler are used for training and testing, respectively. We use AdamW [17] as the optimized method in $1e-5$ learning rate and 0.01 weight decay. The HOLA loss only affects the cross-attention block at $16 \times$ down-sampling rate. In CosmicMan-SDXL, the multi-aspect training strategy [19] in combination with an offset-

noise level of 0.05 is utilized to train our foundation model, and the weight of HOLA loss is set to 1. In CosmicMan-SD, the weight of HOLA loss is set to 0.001 to avoid over-saturation problem. During the training process, a dropout strategy is applied with a probability of 0.1 to all attributes located in front of each object, as well as to the region description phrases, with the notable exception of the phrase “Overall”. Additionally, the photo type, derived from the keypoints, is systematically positioned at the forefront of the dense description. Besides, we also use the quality tuning technique [4] on both models to further enhance the final visually appealing. We finetune on 500 extremely high-quality human images with an offset-noise level of 0.05 for about 1000 iterations. Our model are trained on 32 80G NVIDIA A100 GPUs in a batch size of 64 for about one week.

4.2. Fine-grained Text-Image Alignment Metric

We introduce **Semantic Acc**, a novel text-image alignment metric specifically designed for dense concepts in human images. Firstly, it adopts an atomized approach as other metrics for fine-grained text-image alignment [2, 8], breaking down descriptions into discrete questions to minimize coupling. This atomization is effectively implemented using a predefined question dictionary for finite human descriptions, thereby avoiding the additional errors introduced by generating questions through Large Language Models.



Text Descriptions: An upper body shot, a Caucasian elderly female, natural landscape, white cotton hat, above shoulders wavy white hair, pink floral normal cotton long sleeve shirt

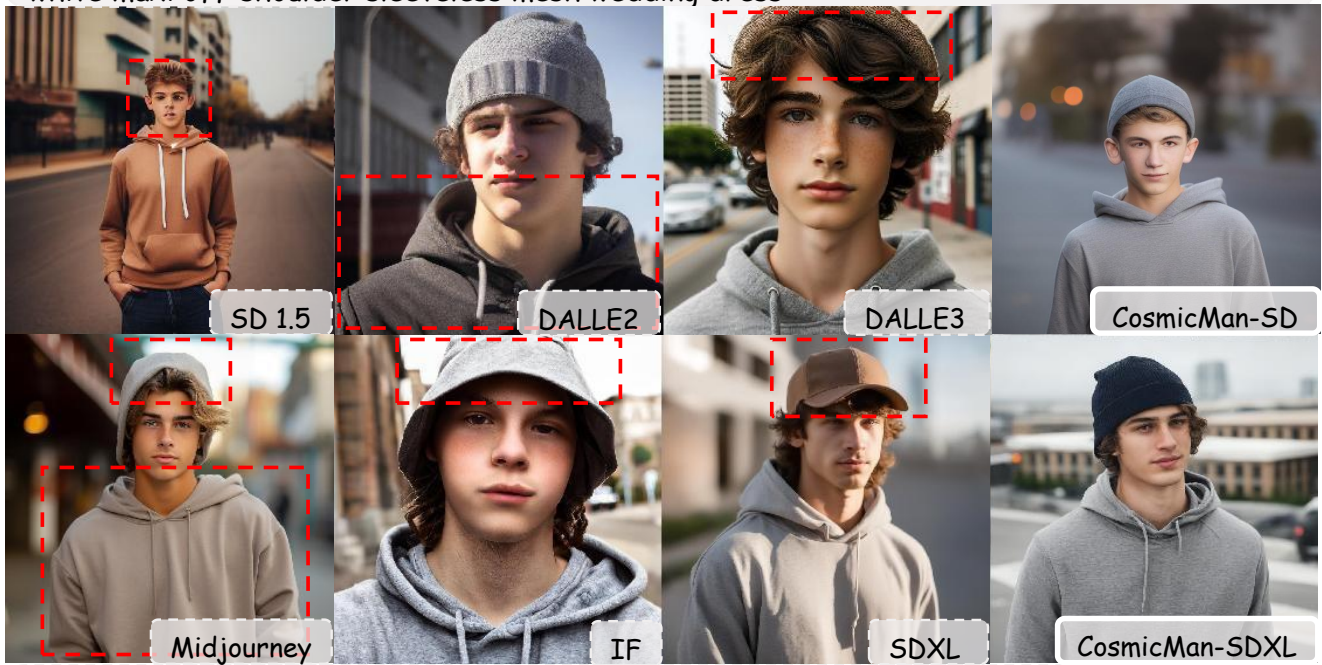


Text Descriptions: A full-body shot, an adult Caucasian female, fit, a white wall, brown straight hair, leather brown mid-calf boots, sleeveless off-shoulder midi silk graphic dress

Figure 6. **Comparison with State-of-the-art Models.** From left to right, the top row features the results of SD 1.5, DALLE2, DALLE3, and CosmicMan-SD. The bottom row presents the results of Midjourney 5.2, DeepFloyd-IF, SDXL, and CosmicMan-SDXL.



Text Descriptions: A nearly full-body shot, an adult Asian female, fit, a forest with trees and a sky with clouds, above eyes bun black hair, solid color cotton black jacket, solid color white maxi off-shoulder sleeveless mesh wedding dress



Text Descriptions: A close up portrait shot, a Caucasian teenager male, fit, a street with a building in the distance, cotton knitted hat, brown wavy above eyes hair, gray cotton long sleeve normal solid color hoodie

Figure 7. Comparison with State-of-the-art Models. From left to right, the top row features the results of SD 1.5, DALLE2, DALLE3 and CosmicMan-SD. The bottom row presents the results of Midjourney 5.2, DeepFloyd-IF, SDXL, and CosmicMan-SDXL.

LAION-5B

HumanSD-1M

CMHQ-1M-AltText

CMHQ-1M-IB

CMHQ-1M-AA

CMHQ-6M-AA



Text Descriptions: A close up portrait shot, an African adult male, fit, rooftop terrace overlooking a sprawling cityscape, crew cut bald black hair, **graphic** blue a jersey normal short sleeve cotton scoop neckline t-shirt, **plastic hat**



Text Descriptions: A close up portrait shot, a Caucasian child female, fit, blurred riverside, normal **graphic** cotton floral short sleeve **pink t-shirt**, close-fitting cotton solid color maxi **gray pants**, brown above shoulders bun hair



Text Descriptions: A full-body shot, an adult Caucasian female, fit, outdoor, solid color black leather **ankle boots high heels**, belt, cotton medium sleeve **plaid short dress**, above chest wavy blonde hair, leather clutch



Text Descriptions: A full-body shot, an Asian adult female, fit, a white wall, short white **stripes cotton skirt**, **solid color** short sleeve cotton **black crop off-shoulder top**, above chest wavy black hair

Figure 8. **Ablation on Training Data.** “AltText” refers to Web Alternative Text, “IB” denotes the image descriptions generated by the pretrained InstructBLIP model, “AA” corresponds to captions produced by Annotate Anyone, and “CMHQ” refers to the CosmicMan-HQ.

For heightened precision, we fine-tuned a Vision-Language Model specifically to answer the atomized questions, rather than relying on pretrained models. Specifically, we ask “yes” or “no” questions for each atomic attribute using InstructBLIP. For example, for a generated image with input description of “plaid short sleeve t-shirt”, we would ask three questions sequentially: “Does the person wear a t-shirt?”, “Does the t-shirt have a plaid pattern?”, and “Is the t-shirt short sleeve?”. The answer to all three questions is “yes”. To prevent overfitting, attributes of the same hierarchy are randomly selected from CosmicMan-HQ 1.0 to generate “no” answer questions during training. The Semantic Acc scores are calculated by dividing the number of correct answers by the total number of questions. To further analyze text-image alignment for humans across various dimensions, we categorize Semantic Acc into three groups: Acc_{obj} for object types, Acc_{tex} for texture attributes, and $\text{Acc}_{\text{shape}}$ for shape attributes, as indexed in Tab. 1.

4.3. Qualitative Comparison to SOTA

We focus on comparing our two foundational models with state-of-the-art models such as SD-1.5, DALLE2, DALLE3, Midjourney, DeepFloyd-IF, and SDXL, as depicted in Fig. 6 and Fig. 7. While SD-1.5 shows the highest CLIPScore among these contenders, as highlighted in the main paper, its visual performance indicates subpar image quality and text-image alignment. Other models are able to generally produce human images that are consistent with detailed descriptions, though they sometimes omit or misinterpret certain concepts, which are highlighted by red dotted boxes in figures. For instance, in the first caption of Fig. 7, which depicts a white wedding dress and a black jacket, these models successfully generate the dress but fail to capture the concept of the black jacket. Regarding CosmicMan-SD, which possesses the same architecture as SD-1.5, it exhibits better text-image alignment. CosmicMan-SDXL excels in both image quality and text-image alignment for dense concepts.

4.4. Qualitative Ablation Studies

In this section, We provide qualitative results based on our setting to show the effectiveness of each part proposed in Daring.

Ablation on Training Dataset. In evaluating the efficacy of data sources, it is observed that the outputs derived from the model trained on CosmicMan-HQ 1.0 dataset exhibit a closer resemblance to authentic imagery and a higher text-image alignment when compared with those generated by models trained on LAION-5B and HumanSD. This comparative analysis is exemplified in Fig. 8, where the first row illustrates a notable discrepancy. The LAION-5B image is characterized by a slight overexposure, while the representation of clothing material in the HumanSD image deviates

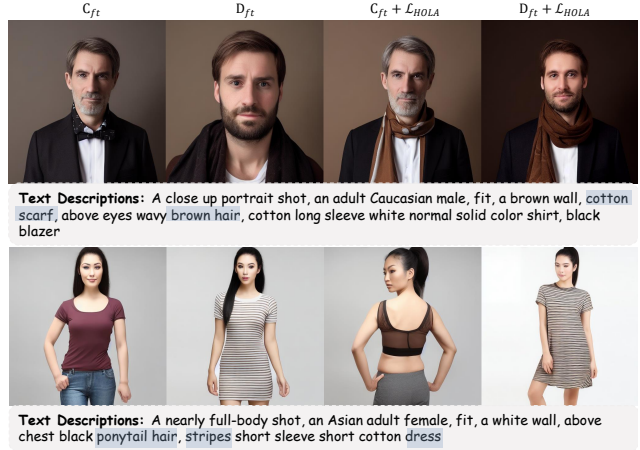


Figure 9. **Ablation on HOLA Loss and Data Discretization.** D_{ft} and C_{ft} denote fine-tuning and testing models on the continuous and decomposed text space. \mathcal{L}_{HOLA} represents the human body and outfit-guided loss for alignment. The blue backgrounds are utilized to accentuate the content of the generated pairs. This approach is consistently applied throughout the following images to maintain uniformity and enhance visual clarity.

from a cotton-like texture.

Regarding the impact of data scaling on model performance, the training on the CosmicMan-HQ-6M dataset demonstrates enhanced precision in the text-image alignment for dense concepts. As evidenced in Fig. 8, the model trained with CosmicMan-HQ-6M more accurately captures fine-grained attributes, such as ankle boots (illustrated in the third row) and a striped skirt (depicted in the fourth row), in contrast to its CosmicMan-HQ-1M counterpart.

Furthermore, the superiority of annotation quality in the Annotate Anyone becomes evident when contrasted with the pretrained IB model. This is particularly noticeable in the generation of more accurate and refined annotations, leading to significant improvements in the visual representation of items such as a graphic pink t-shirt and gray pants, as showcased in the second row of Fig. 8.

Ablation on Training Strategy. We delineate the impact of Data Discretization for Humans and the incorporation of HOLA loss on the accuracy of the generated images. Fig. 9 elucidates this effect by contrasting results obtained under decomposed and continuous text spaces, underscoring the significance of Data Discretization for Humans. The model D_{ft} demonstrates a heightened ability to accurately generate images of a cotton scarf (as depicted in the first row) and a striped dress (shown in the second row), compared to its counterpart, C_{ft} . This observation underscores the premise that a discrete text space is more adept at handling complex, dense concepts in our dataset. Nevertheless, the absence of constraints on key K in the attention block occasionally leads to instances of attribute misalignment. In



Figure 10. **Ablation on Terms of HOLA Loss.** Baseline means only fine-tuning the model on CosmicMan-HQ 1.0. Term 1 refers to the first term of HOLA loss, which works under the guidance of human body structure. Term 2 refers to the second term of HOLA loss, which helps reduce the ambiguities of outfit-level descriptions.

an effort to address this, the integration of HOLA loss with Data Discretization for Humans is explored. This combination, represented as $D_{ft} + \mathcal{L}_{HOLA}$, culminates in a further enhancement of semantic alignment. The resulting images depict a more realistic representation of a cotton scarf and accurately colored hair in the first row, and a more precisely rendered striped dress and ponytail hair in the second row.

In an effort to further dissect the influence of HOLA loss, a comparative analysis focusing on the individual contributions of each term within the loss function is conducted. This is depicted in Fig. 10, where the differential impacts of these terms are illustrated. Specifically, Term 1 is observed to facilitate the improved generation of attributes that are characteristic of larger regions. Examples of this can be seen in the enhanced depiction of a blue shirt and a black skirt on the left and right images, respectively. However, it is noted that Term 1 is less effective in accurately rendering attributes that are localized. For instance, the polo shirt on the left side is more akin to a t-shirt in its representation. This discrepancy can be attributed to the nuanced differences, such as the localized collar shape, which distinguishes a polo shirt from a t-shirt. By composing Term 2, the model exhibits a heightened proficiency in generating attributes pertinent to localized regions. This is evident in the more accurate portrayal of the cat pattern and the polo shirt on the left, as well as the distinct pattern of the t-shirt on the right.

4.5. Qualitative Results of Different Granularity

To illustrate CosmicMan’s ability in processing descriptions of varying granularity, we offer a visualization in Fig. 11 that features captions with progressively increasing detail. Initially, we start with a simple caption specifying different outfit types to generate the base human image. Subsequently, we incrementally enrich the description by adding details about texture, shape, and color. The step-by-step result reveals our model’s capability to produce high-quality human images that remain faithfully aligned with increas-

ingly complex and dense concepts.

4.6. Qualitative Results of Different Concepts

We demonstrate the versatility of CosmicMan-SDXL in handling varying fine-grained descriptions, as depicted in Fig. 12. We only modify certain elements of the descriptions while keeping other components consistent. In the first row, we alter the descriptions of outfit shape, such as the sleeve length in the target region, resulting in visible changes, yet the overall spatial layout is maintained. Next, we explore different textures and colors, like fur floral dress, showcasing the model’s ability to incorporate unique dense concepts. Lastly, when modifying the outfit type, significant changes are observed in the human structure and scene layout.

4.7. Quantitative Comparison on Unseen Testset

We follow the experimental settings and provide an additional zero-shot evaluation on an unseen human subset, which contains 2048 images filtered from the MS-COCO 2014 validation subset [15]. As shown in Tab. 5, our models outperform all other methods in terms of image quality (FID). Regarding text-image alignment, we compare CLIPScore instead of Semantic Acc since the captions are in free-form text format. Its results consistently align with those from our original test set.

Table 5. **Quantitative Comparison on Unseen Human Testset.** We conduct zero-shot evaluation on MS-COCO 2014 validation subset. “CM-SD” and “CM-SDXL” are short for CosmicMan-SD and CosmicMan-SDXL.

Metrics	SD 1.5	SD 2.0	SDXL	DF-IF	CM-SD	CM-SDXL
FID	50.45	75.71	50.16	53.38	49.34	45.90
CLIP	27.23	23.58	27.37	26.96	27.23	25.56
Speed (ms/f)	43.66	55.99	56.23	203.43	43.53	56.32



Figure 11. **Qualitative Results of CosmicMan-SDXL with Increasing Granularity.** A simple description is initially provided, from which the granularity is progressively increased, emphasizing texture, shape, and color, leading to the generation of the following human images.

4.8. Applications

In this subsection, we provide more qualitative results on two downstream applications: 2D human editing and 3D human reconstruction to show the effectiveness of our specialized foundation model.

2D Human Editing. We compare our ComicMan-SDXL with SDXL pretrained model using T2I-Adapter [18]. We use skeleton maps extracted by Openpose [1] as guidance to generate portraits with specified poses. The first and second rows in Fig. 13 show that our results exhibit more accurate pose control. Besides, SDXL tends to generate semantic inconsistency images with a hazy background, while our method generates more realistic and semantic-consistent images.

3D Human Reconstruction. We further compare our ComicMan-SD with SD-1.5 pretrained model based Magic123 [20] for 3D human reconstruction. The first two examples in Fig. 14 show the ability on Image-to-3D, where both reference image and prompt are used as input. The first example shows that our model can maintain the hat shape of the girl in each view, demonstrating that our model possesses better multi-view consistency. The second example shows Stable Diffusion pretrained model tends to generate results with vague and odd human shapes (the 3rd and 6th images of SD in second example). In contrast, our results obtain more accurate human body and face geometric shapes. The third example in Fig. 14 shows the ability on Text-to-3D, where the single prompt is initially input to CosmicMan-SD or Stable Diffusion pretrained model to generate the reference image, and then sent to Magic123. It can be seen that our results are significantly superior to

Stable Diffusion results on both text-image consistency and geometric shape.

4.9. More Generated Samples of CosmicMan

In Fig. 15 ~ Fig. 20, we present additional generated samples showcasing a variety of captions and aspect ratios. Please note that some images have been cropped for improved typesetting.



Shape: short sleeve normal blouse, straight jeans



Shape: medium sleeve crop blouse, close-fitting jeans



Shape: long sleeve v neckline blouse, wide-leg jeans



Texture and color: yellow cotton graphic dress



Texture and color: white lace dress



Texture and color: fur floral dress



Outfit type: t-shirt, skirts, ankle boots, beret hat



Outfit type: coat, leggings, sneakers, baseball cap

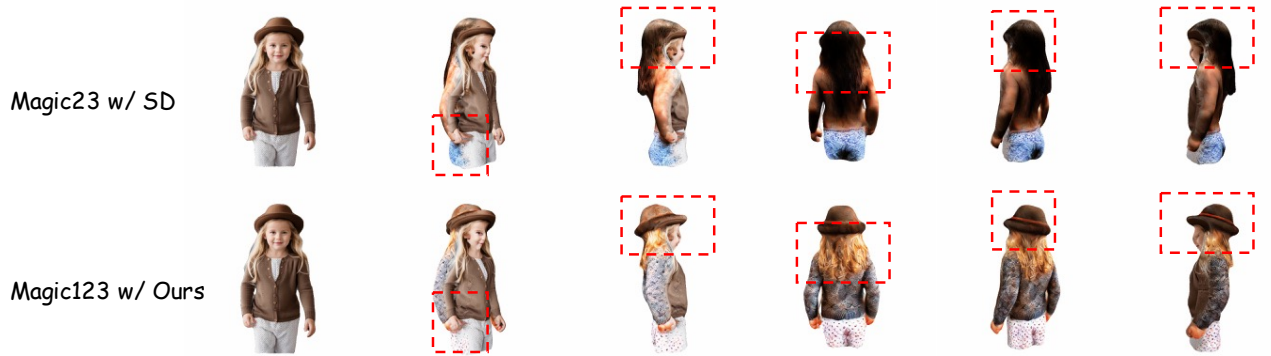


Outfit type: cardigan, sweaters, pants, high heels, cowboy hat

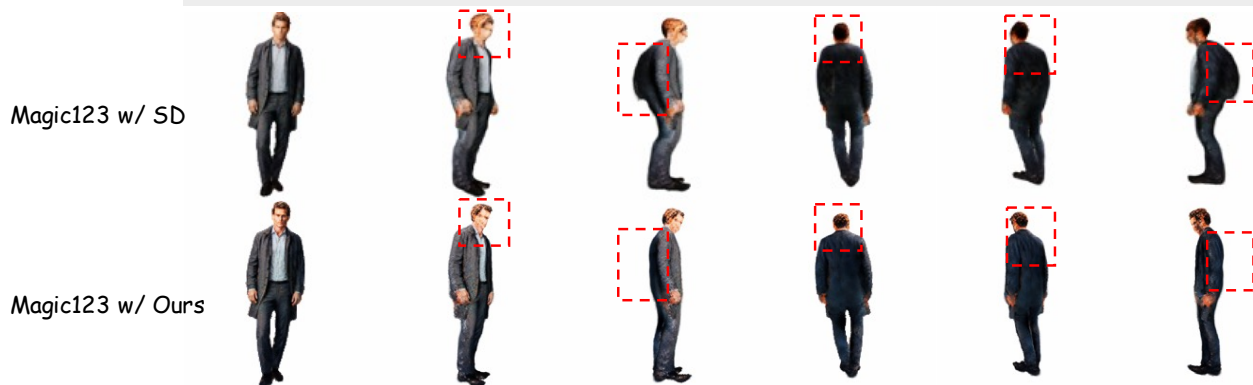
Figure 12. Qualitative Results of CosmicMan-SDXL with Different Descriptions on Shape, Texture, and Outfit Types.



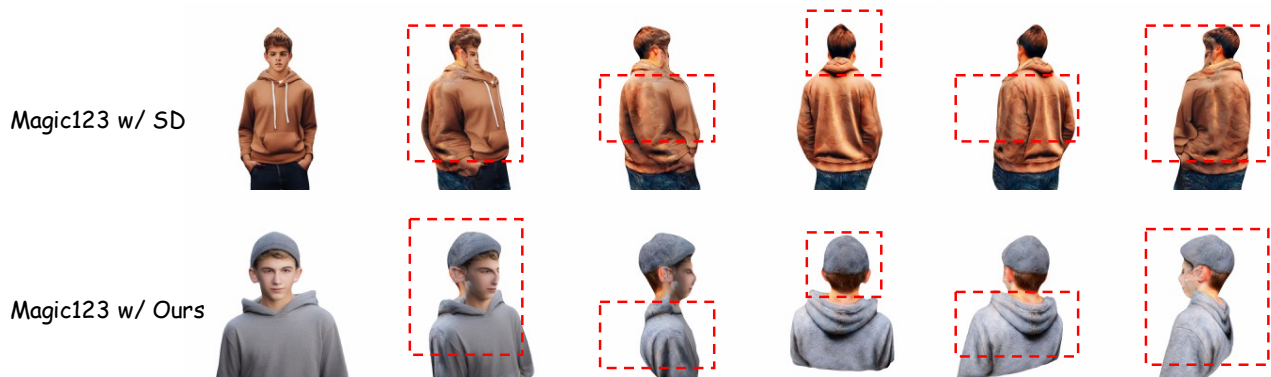
Figure 13. **Visualization of 2D Human Editing.** We compare our CosmicMan-SDXL with SDXL pretrained model based on T2I-Adapter [18]. We highlight the areas of text-image inconsistency of SDXL results in “Text Descriptions” using blue background.



Text Descriptions: An upper body shot, a Caucasian female child, fit, above chest blonde wavy hair, woven hat, straight maxi cotton graphic white polka dots pants, solid color normal woven brown long sleeve sweaters



Text Descriptions: A full-body shot, an adult Caucasian male, fit, wavy gold above eyes hair, solid color cotton normal long sleeve shirt, cotton solid color overcoat, solid color cotton straight maxi pants, solid color shoes



Text Descriptions: A close up portrait shot, a Caucasian teenager male, fit, a street with a building in the distance, cotton knitted hat, brown wavy above eyes hair, gray cotton long sleeve normal solid color hoodie

Figure 14. **Visualization of 3D Human Reconstruction.** We compare our CosmicMan-SD with Stable Diffusion pretrained model based on Magic123 [20]. The first and second examples show the ability of Image-to-3D. The third example shows the ability of Text-to-3D. CosmicMan-SD can generate results with better multi-view consistency and geometric shapes. The first image in each row is the reference image. Note that the reference images in the fifth and sixth are generated by CosmicMan-SD and SD-1.5 respectively. The red box highlights the improvement of our foundation model compared to Stable Diffusion.



Text Descriptions:

A close up portrait shot, an adult Caucasian female, fit, runway, cotton long sleeve solid color white normal t-shirt, solid color fur scarf, wavy above chest brown hair.



Text Descriptions:

A headshot, an adult Latino male, fit, softly blurred city traffic at night, solid color maxi cotton dress, above eyes wavy brown hair, solid color long sleeve cotton normal red t-shirt, cotton baseball cap.

Figure 15. More Generated Samples of CosmicMan.



Text Descriptions:

A close up portrait shot, an Asian child female, fit, fuzzy urban park with blurred walkers, sleeveless white cotton solid color normal camisole, above shoulders black bob hair.



Text Descriptions:

A close up portrait shot, an African adult female, fit, autumn forest, warm earthy blur, printed cotton scarf, cotton long sleeve black normal solid color t-shirt, above shoulders afro-hair black hair.

Figure 16. More Generated Samples of CosmicMan.



Text Descriptions:
A headshot, an adult Caucasian female, misty forest with distant mountains, above eyes brown bob hair.



Text Descriptions:
A headshot, a Caucasian adult male, fit, outdoor, normal white solid color long sleeve cotton shirt, white bald hair, solid color black jacket.



Text Descriptions:
A headshot, an Asian adult female, fit, a mountainous landscape, solid color cotton long sleeve short dress, above chest brown straight hair, solid color gray cotton overcoat.

Figure 17. More Generated samples of CosmicMan.



Text Descriptions:
A headshot, an adult Caucasian female, fit, softly blurred city traffic at night, above chest brown straight hair, gray cotton normal long sleeve solid color sweaters.



Text Descriptions:
A close up portrait shot, an adult Caucasian male, vaguely visible park with autumn trees, above eyes wavy brown hair, cotton long sleeve white normal solid color shirt.

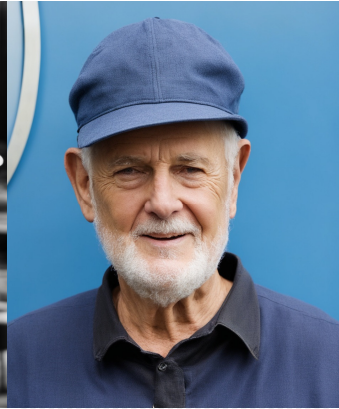


Text Descriptions:
A close up portrait shot, an adult Caucasian female, fit, blurred riverside with distant boats, wavy red above chest hair, blue silk long sleeve normal blouse, silk midi long sleeve dress.

Figure 18. More Generated Samples of CosmicMan.



Text Descriptions:
A headshot, an adult Caucasian female, black and white city in distance, above chest blonde wavy hair.



Text Descriptions:
A headshot, a Caucasian elderly male, a blue wall, bald above eyes gray hair, blue short sleeve polo shirt, blue hat.



Text Descriptions:
A full-body shot, a Caucasian adult female, fit, outdoor, sleeveless solid color silk short off-shoulder jumpsuit, black bun above shoulders hair, leather solid color high heels



Text Descriptions:
An upper body shot, an adult Asian female, fit, dusk cityscape in dreamy blur, grey tie-dye long sleeve graphic cotton normal sweatshirt, tie-dye cotton graphic tapered grey maxi sweatpants, wavy above chest brown hair.



Text Descriptions:
An upper body shot, a Caucasian elderly female, fit, in a city traffic scene, white cotton hat, above shoulders wavy white hair, pink floral normal cotton long sleeve shirt.

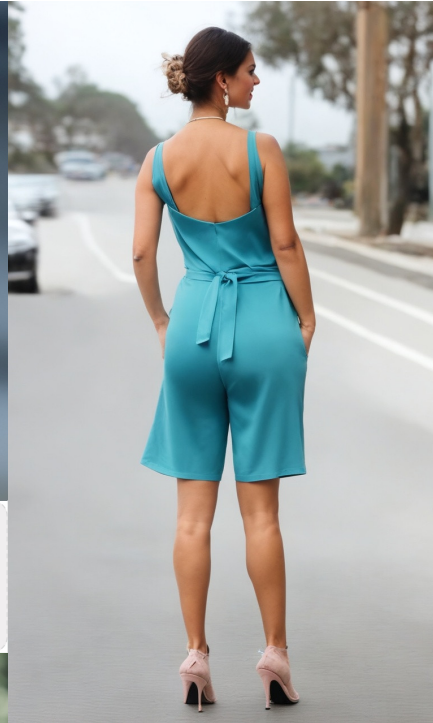
Figure 19. More Generated Samples of CosmicMan.



Text Descriptions:
A close up portrait shot, a Caucasian child female, fit, a white wall, normal graphic cotton floral short sleeve gray t-shirt, pink solid color leather sneakers, close-fitting cotton solid color maxi pink pants, brown above shoulders bun hair



Text Descriptions:
A close up portrait shot, a Caucasian toddler male, fit, outdoor, blonde straight above eyes hair, long sleeve normal blue solid color cotton t-shirt.



Text Descriptions:
A full-body shot, a Caucasian adult female, fit, outdoor, sleeveless solid color silk short jumpsuit, black bun above shoulders hair, solid color pink high heels.



Text Descriptions:
A headshot, a Caucasian adult male, fit, summer forest, cold earthy blur, normal black solid color long sleeve cotton shirt, white bald hair, black ankle boots leather solid color loafers.



Text Descriptions:
An upper body shot, a Caucasian adult female, fit, blurred riverside, fur solid color scarf, blonde wavy above shoulders hair, gray normal silk long sleeve solid color blouse, white solid color maxi straight cotton pants.



Text Descriptions:
A close up portrait shot, an adult Caucasian female, fit, indoor, short sleeve solid color orange silk dress, leather belt, above shoulders wavy blonde hair.

Figure 20. More Generated Samples of CosmicMan.

References

- [1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. **13**
- [2] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. **7**
- [3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. **2, 4**
- [4] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jiali Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. **7**
- [5] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. **1**
- [6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. **7**
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. **7**
- [8] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. **7**
- [9] Tanuj Jain, Christopher Lennan, Zubin John, and Dat Tran. Imagededup. <https://github.com/ideal0/imagededup>, 2019. **2**
- [10] 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 *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 316–332. Springer, 2020. **2**
- [11] Byunggho Jo, Donghyeon Cho, In Kyu Park, and Sungeun Hong. Ifqa: Interpretable face quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3444–3453, 2023. **2**
- [12] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 618–629, 2023. **2**
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. **4**
- [14] Baijiong Lin, Feiyang Ye, Yu Zhang, and Ivor W Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *arXiv preprint arXiv:2111.10603*, 2021. **6**
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV 2014*, 2014. **12**
- [16] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. **1, 2**
- [17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. **7**
- [18] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongqiang Qi, Ying Shan, and Xiaoohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. **13, 15**
- [19] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. **7**
- [20] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. **13, 16**
- [21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. **7**
- [22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. **1**
- [23] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. **2**
- [24] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. **2**
- [25] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. **2**

- [26] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023.