CosmicMan Prime: Enhancing Activation Strategies for Improved Text-to-Image Synthesis

Gabriel Odai Afotey

April 2025

Abstract

I present **CosmicMan Prime**, a text-to-image foundation model designed to enhance image generation quality through refined activation strategies. Unlike traditional models that struggle with suboptimal image quality and text-image misalignment, CosmicMan Prime achieves improved synthesis by optimizing activation functions, ensuring better alignment between textual descriptions and generated images.

Central to CosmicMan Prime's improvements are the modified activation strategies and data-driven refinements: (1) I investigate the impact of activation functions on model performance, proposing a shift from the standard SiLU activation to more robust alternatives. By testing various activation functions, I demonstrate that carefully chosen activations can significantly enhance model convergence and output fidelity. This exploration provides a deeper understanding of how activation functions influence the relationship between text and image data. (2) I argue that a specialized text-to-image model must balance both architectural improvements and practical usability. To this end, we focus on optimizing CosmicMan Prime's underlying structure to enhance its ability to generate high-quality images from detailed textual descriptions. This model integrates the improved activation function strategies into existing architectures, ensuring scalability and effectiveness without requiring substantial changes to the foundational model.

Through these refinements, CosmicMan Prime's delivers higher-quality, more realistic images, while maintaining the flexibility to be used across various downstream tasks. Project Link: https://github.com/odai307/CosmicMan-Prime

1 Introduction

Text-to-image foundation models, such as Stable Diffusion (SD) [43], Imagen [45], and DALLE [40], have revolutionized the field of Computer Vision and Graphics by generating high-quality images from textual descriptions. These models, powered by large-scale image-text datasets [47, 3] and advanced generative algorithms [51, 11, 20], exhibit impressive abilities in creating images with remarkable detail. Underpinned by robust prior knowledge, these models have significantly contributed to downstream applications, such as DreamBooth [44] and ControlNet [58] for 2D image generation, and DreamFusion [38] and Zero-1-to-3 [30] for 3D object creation. However, a critical gap persists in the generation of *human-centric* content, specifically the lack of a specialized text-to-image model tailored for human subjects.

Previous research has focused on human-centric content generation tasks, including 2D human generation/editing [14, 22, 28] and 3D human generation/reconstruction [56, 46, 57, 16]. However, these approaches have often faced limitations due to narrow datasets [6], biased distributions [14], or lack of quality [25]. Achieving generalization across a wide range of human identities, appearances, and geometries in real-world scenarios has been challenging. Despite these challenges, the rise of general-purpose text-to-image models has opened up new opportunities for revolutionizing human-centric content generation by offering enhanced generalization capabilities.

The critical question now is: *How can we create a specialized text-to-image foundation model for humans?* We identify three essential components for such a model: 1) **High-Quality Data.** The quality of the raw data plays a vital role in training a successful foundation model. This includes not only the volume of data but also the quality and diversity of images, as well as the precision and comprehensiveness of

annotations. While datasets such as LAION-5B [47] and COYO-700M [3] have driven advancements in general-purpose models, they often fail to accurately capture the diversity of human forms and contain noise in the annotations. 2) **Scalable Data Production.** A successful human-specialized model requires the ability to scale with the growth of real-world data. Developing a cost-effective, scalable data production process is crucial. Traditional annotation methods are often costly and slow [29, 48, 1, 36], and static datasets struggle to adapt to evolving real-world data distributions. 3) **Pragmatic Model.** The model must be simple to integrate into downstream tasks and capable of generating high-quality outputs with realistic human structures and precise text-image alignment. Despite existing models like MidJourney [33], DALLE [40], and SD [43], challenges remain in achieving high-fidelity human generation that meets the complex requirements of human anatomy.

We introduce **CosmicMan+**, a specialized text-to-image foundation model tailored for human content generation. The core innovation of CosmicMan+ lies in a refined approach to activation functions, which enhances model performance in generating more realistic human images. By shifting from traditional activation functions like SiLU to more robust alternatives, we significantly improve the synthesis quality and text-image alignment.

We also propose a new data production paradigm, *Annotate Anyone*, powered by human-AI collaboration. This paradigm continuously produces high-quality, cost-effective data by sourcing images from diverse academic datasets and the internet, and iteratively refining annotations with human-in-the-loop techniques. This dynamic data flywheel supports the rapid growth of a high-quality, human-centric dataset, *CosmicMan-HQ 1.0*, which currently includes 6 million human images with a mean resolution of 1488×1255, and 115 million attributes, texts, and annotations.

Building on CosmicMan-HQ, we present a human-specialized foundation model that supports various content generation tasks. The model leverages minimal modifications to SD's architecture, focusing on enhanced activation strategies and integration into downstream tasks. Additionally, we introduce the *Decomposed Attention Refocusing* (Daring) training framework, which decomposes cross-attention features into groups based on human body structure. This method improves text-image alignment and generates high-quality outputs with realistic human forms. A new loss function, *HOLA* (Human Body and Outfit Guided Loss for Alignment), supervises the model to improve attention focusing.

Our experiments demonstrate that CosmicMan+ achieves superior image quality and text-image alignment compared to existing state-of-the-art models. We conduct extensive ablation studies to validate the effectiveness of our approach in data production and model training. Finally, we showcase the versatility of our model through applications in both 2D and 3D human content generation.

2 Related Work

2.1 Text-to-Image Foundation Models

The evolution of text-to-image foundation models has led to the generation of high-fidelity images that adhere to text descriptions. DALLE [40], the pioneering model in zero-shot text-to-image generation, autoregressively models text and image tokens within a unified data stream. Its successors [41, 2] have improved performance by refining model architecture and enhancing captions. Imagen [45] employs a larger text encoder that boosts photo-realism in generated images. Open-source models such as DeepFloyd-IF [9], PixelArt-α [5], and particularly SD [43] and SDXL [37] have played a pivotal role in driving widespread adoption, fostering numerous applications across downstream tasks. Innovations like ControlNet [58] and T2I-Adapter [34] have advanced 2D content generation, while 3D models like Zero-1-to-3 [30] and DreamFusion [38] leverage SD to generate high-quality 3D objects.

Despite these advances, existing foundation models are designed for general-purpose content generation and often fall short in producing realistic human representations. These models tend to overlook the subtleties of human anatomy and clothing, posing a challenge for human-centric content generation. Thus, there remains a gap for a specialized text-to-image foundation model tailored specifically to human content.

2.2 Text-Driven Human Image Generation

Previous research has focused on human image generation in specific domains, particularly fashion, with notable successes in both generation and editing tasks. For example, Text2Human [22] uses a two-stage

framework with VQ-VAE [51] to transform human pose into human parsing, incorporating texture descriptions for human image generation. FashionTex [28] enables text and texture-based control for virtual try-on by utilizing pretrained generative models [14]. However, these methods are often limited by biased and insufficient training data, which restricts the diversity of generated images. Approaches such as HumanSD [25] and ControlNet [58] leverage foundation models like SD to generate more diverse human images by adding additional conditions, such as skeletons and normal maps. Despite these advancements, these models are still focused on controlled generation and do not provide the flexibility needed for diverse human image generation.

In contrast, CosmicMan stands out as a specialized foundation model designed to produce high-quality, diverse human images without relying on spatial conditions during inference. This enables a broader range of human-centric applications without being constrained by domain-specific limitations.

2.3 Text-Image Alignment for Dense Concepts

Early text-to-image models, which were primarily trained on short captions, struggled to capture the complexity of dense concepts described in longer, more detailed text. Dense concepts, as explored in various text-to-image benchmarks [21, 15], involve multiple objects, attributes, and spatial relationships that describe an image from different perspectives. The challenge lies not only in generating each element but also in accurately depicting their relationships within long descriptions.

Recent research has highlighted the critical role of cross-attention mechanisms in text-to-image alignment. Studies such as Prompt-to-Prompt [17], Attend-and-Excite [4], Structurediffusion [12], and others [53, 27, 42] have shown that utilizing gradients from cross-attention maps can improve latent feature alignment during the diffusion process. Additionally, FastComposer [55] supervises cross-attention maps to refine the alignment during training.

In the context of human image generation, directly applying these methods becomes more complex due to the dense nature of descriptions associated with human images. Captions for human images often involve multiple attributes clustered in a small image region. For example, our dataset contains an average of 30 dense attribute descriptions per person, with 5 concepts associated with each sub-body region (as shown in Fig. 5). To address this challenge, we propose a training framework that leverages human-specific priors to decompose the text-image data and supervise cross-attention maps, enhancing the alignment of dense concepts in human-centric content generation.

3 Annotate Anyone – A Data Flywheel

To enable the learning of the human-specialized foundation model, we propose a human-AI cooperation paradigm for data production named **Annotate Anyone**. It combines the strengths of AI and human expertise to build a continuously expandable dataset, CosmicMan-HQ, with rich annotations.

In this section, we will first introduce Annotate Anyone by comparing it to previous data production paradigms (Sec. 3.1). Then, we will elaborate on the procedures of Annotate Anyone in data sourcing and annotation (Sec. 3.2). Finally, we will analyze the statistics of the CosmicMan-HQ 1.0 dataset produced by Annotate Anyone and show its superiority to existing datasets (Sec. 3.3).

3.1 Data Production by Human-AI Cooperation

To construct large-scale image datasets with labels, there are mainly two paradigms – by humans or by AI. Data production by humans (as depicted in Fig. 1 (a)) needs human annotators to manually label images one by one [10, 29, 60], which suffers from its high cost and thus is hard to scale up to support the recent development of large foundation models. On the other hand, data production by AI (as depicted in Fig. 1 (b)) uses off-the-shelf models to get labels for free [2, 5]. Although this paradigm dramatically reduces costs and is easy to scale up, it is notorious for its noisy, jagged, and coarse labeling results. Moreover, both of these paradigms rely on fixed datasets for labeling, which results in limited diversity and severe bias versus real-world data. To train a large foundation model, a huge quantity of data, high-precise and fine-grained labeling, and real-world distribution are all indispensable. Thus, these paradigms are especially knotty to adapt to the human domain.

To this end, we propose a new data production paradigm by *human-AI cooperation* named Annotate Anyone (as shown in Fig. 1 (c)). Compared to data production by humans and AI, Annotate Anyone pivots

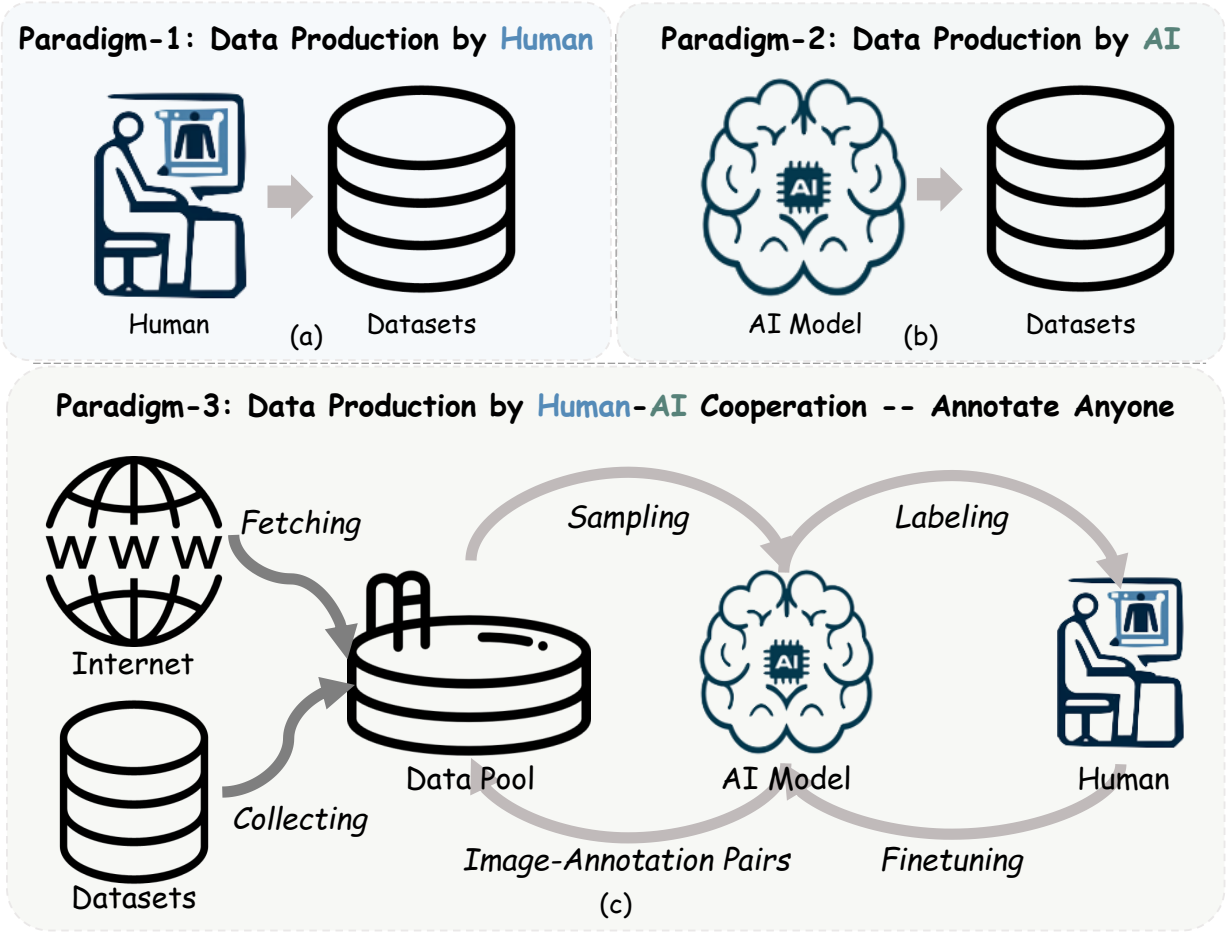

Figure 1: **Data Production Paradigm.** (a) Data production by humans and (b) data production by AI. (c) Our proposed new data production paradigm by Human-AI cooperation, named Annotate Anyone. It serves as a *data flywheel* to produce dynamic up-to-date high-quality data at a low cost.

on two characteristics: 1) flowing data, and 2) human-in-the-loop annotation. Flowing data is sourced from two origins: existing published datasets and the Internet. By collecting data from existing published datasets, such as SHHQ [14] and LAION-5B [47], we can upcycle them to match the qualifications of a high-quality human dataset. By fetching data from the Internet, we can obtain the massive data produced by human beings every second. Our data sourcing system is always on call to run when the data quantity triggers the lower-bound threshold. Thus, our data pool is continuously *flowing and refreshed*, which distinguishes it from previous paradigms. Human-in-the-loop annotation coordinates three entities: data pool, AI, and human annotators, to work in a circle. By labeling a small quantity of data with the greatest necessity, human annotators and AI models cooperate to iteratively improve the quality of data annotation. Consequently, data in the pool will have progressively better annotation quality at a minimal cost.

With Annotate Anyone, we construct a **data flywheel** to enable a dynamic up-to-date production of high-quality data.

3.2 Procedure of Annotate Anyone

The procedure of Annotate Anyone consists of two main parts: 1) flowing data sourcing to obtain and filter high-quality data from both academic datasets and the Internet continually, and 2) human-in-the-loop data annotation to get precise yet cost-effective labels covering detailed dense concepts and ensuring high-quality annotations.

3.2.1 Flowing Data Sourcing

We first source images from various origins to ensure massive quantity and catch the real-world distribution. Then, we design a data filter to eliminate unbefitting images for human content generation tasks.

Data Origins. We start with three *academic* datasets to recycle existing data resources: LAION-5B [47], SHHQ [14], and DeepFashion [31]. LAION-5B is a renowned collection of massive images shared online, while the other two datasets are smaller in scale and diversity but meticulously curated to ensure high quality. Then, we initiate 128 parallel processes in 32 CPU servers, monitoring a wide spectrum of APIs on the Internet, including Flickr [13], Unsplash [50], Pixabay [35], . These APIs give access to a vast collection of growing and diverse images, rendering a real-world distribution.

Data Filtering. The current data pool exhibits a broad distribution, but high-resolution human images are not the primary constituent. We use a set of data filtering strategies to distill a high-quality human-centric subset, including fake-people detection, image quality assessment, and so on. To remove fake-people images (, cartoon characters, mannequin models, and generated images), we fine-tuned Eva-CLIP [49] with Human-Art [24] 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 [59], IFQA [23], and HPSv2 [54]) are applied, estimating the image quality on both face and global levels, to streamline the data pool further. Further, we utilize YOLOv7 [52] to filter out images without humans or those containing more than one individual. Images where the largest detected face is smaller than 224×224 or where the image is smaller than 640×1280 are removed as well. See more details in the supplementary material.

3.2.2 Human-in-the-loop Data Annotation

Having a data pool with diverse and high-quality human images, the next step is to possess precise, fine-grained yet cost-effective annotations for the images. We propose a human-in-the-loop data annotation workflow to iteratively refine the labeling quality of the data in the pool. Below, we first introduce the annotation iteration and then discuss the label protocol we use to describe humans in a detailed way.

Annotation Iteration. As shown in Fig. 1 (c), the iterations start from sampling an image set I_i from the data pool and end up with putting all image-annotation pairs (I, A) back to the data pool. We set an evaluation set I_e with ground truth. In each iteration, I_e is used to determine the categories that need to be labeled by human annotators, and I_i will be partially labeled with the selected categories. Then, I_i is used to finetune the AI model. The finetuned AI model is evaluated on I_e to determine whether to continue or stop the iterations. Finally, a well-finetuned AI model is used to get image-annotation pair (I, A) for all data in the pool. Specifically, inspired by methods [2, 5] that use the Vision-Language Model (VLM) to perform image captioning tasks, we leverage a pretrained InstructBLIP [8] as our AI model in the iteration. Please refer to the supplementary for the pseudo-code of the annotation iteration.

Figure 2: **Parsing Examples from CosmicMan-HQ.** The parsing results of sampled image in our dataset, along with detailed labels for each part. Text descriptions are obtained from labels.

The pivotal mechanism to implement the high-precise yet low-cost annotation is the trigger of human annotation. During the initial iteration, the annotation team labels all categories based on 70 questions. We observed that the accuracy of the predicted labels follows a real-world distribution, exhibiting a long-tail distribution. For the head categories, such as age and gender, the pretrained AI model already proficients in the prediction. Thus, in subsequent iterations, human annotators focus on tail categories, and categories with an accuracy above 85% will no longer be manually labeled. Our iterative process significantly improved the

Table 1: **Dataset Comparison.** The statistical comparison between publicly available human-related datasets and CosmicMan-HQ 1.0. "Common Scale" refers to the dataset that includes images captured at common scales, such as full-body shots, portrait photos, and half-body shots. "HP" and "Aes" refer to Human Parsing maps and Aesthetic scores.

	Data Quantity			Imaging Quality		Annotation							3*Domain
	Total Image #	Mean Resolution	Common Scale	Global↑	Face↑	Cat #	Attr #	Text	Bbox	Kpts	HP	Aes	
Human-Art [24]	$50K$	1115×1287	[HTML]A5B592	3.42	2.87	-	-	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	[HTML]D092A7	[HTML]D092A7	Real world & AI
DF-MM [22]	$44K$	750×1101	[HTML]D092A7	4.64	3.38	18	$587K$	[HTML]A5B592	[HTML]A5B592	[HTML]D092A7	[HTML]A5B592	[HTML]D092A7	Fashion
LAION-Human [25]	$1M$	688×650	[HTML]A5B592	4.20	2.66	-	-	[HTML]A5B592	[HTML]D092A7	[HTML]D092A7	[HTML]D092A7	[HTML]A5B592	Real world
SHHQ 1.0 [14]	$40K$	1024×512	[HTML]D092A7	4.23	2.13	-	-	[HTML]D092A7	[HTML]D092A7	[HTML]A5B592	[HTML]A5B592	[HTML]D092A7	Fashion
CosmicMan-HQ 1.0	$6M$	1488×1255	[HTML]A5B592	4.37	3.37	70	$115M$	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	Real world

VLM model's overall accuracy by at least 30% compared to the pretrained model. Moreover, the progressive reduction of labeled annotations during the iterations resulted in only 1% compared to full manual labeling. Please refer to the supplementary materials for experimental details.

Label Protocol. To systematically describe a human image comprehensively, we design a label protocol that leverages the human parsing model SCHP [26] to break down an image into 18 fine parts, including background, face, upper clothing, and . Each part has an average of 3 to 8 associated questions, resulting in a total of 70 questions that correspond to 70 categories. For example, "top-sleeve length" is one of the categories. The answers under each category form the detailed attributes in our dataset. For instance, the attributes associated with "top - sleeve length" include "long sleeve, mid-sleeve, sleeveless", and so on. As shown in Fig. 5, for every single clothing item, once we confirm the existence of a particular garment, we proceed to further inquire about deeper detailed questions, such as color, pattern, and material. Questions related to global attributes are also asked. By employing this label protocol, we obtain a rich set of attributes regarding the human image with each label corresponding to a specific spatial position. Please refer to supplementary materials to see our final labeling results with the comparison to others.

3.3 CosmicMan-HQ 1.0 Dataset

By running Annotate Anyone, so far, the first version of the produced dataset CosmicMan-HQ 1.0 consists of $6M$ high-resolution, single-person images, along with corresponding rich annotations. Here we compare our dataset with representative human-centric datasets in terms of data quantity, imaging quality, and annotations.

As depicted in Tab. 3, our dataset is the largest crafted human-centric dataset, six times larger than LAION-Human [25]. The mean resolution is 1488×1255, surpassing previous human-only datasets like DF-MM [22] and SHHQ [14] by a large margin. Our dataset possesses a diverse collection of human images, including full-body shots, headshots, half-body shots, and so on. In terms of image quality at both overall and face level, our dataset ranks second only to the fashion-focused DF-MM dataset, which predominantly contains professional studio images but with less diversity and a data amount. As for annotation, only DF-MM and ours provide manually labeled categories, but the former dataset is much smaller in data volume and category numbers. CosmicMan-HQ 1.0 provides 70 categories and around $115M$ detailed attributes as detailed in the Label Protocol part of Sec. 3.2.2.

Highlighting our dataset's uniqueness, CosmicMan-HQ 1.0 distinguishes itself by providing an unparalleled wealth of diverse annotations, including $115M$ attributes, texts, bounding boxes, keypoints, human parsings, and rich meta information (web alternative texts, aesthetic scores, watermark scores, face/global quality scores, and camera EXIF parameters).

4 Daring - The Training Framework

We propose **Daring** (**D**ecomposed-**A**ttention-**R**efocus**ing**), a training framework rooted in original Stable Diffusion (SD) with minimal modification. The framework is illustrated in Fig. 6. It enjoys three properties at the same time – friendly to computational costs, compatible with downstream tasks supported by SD, and robust in producing high-quality human images that align well with dense concepts. These come from two parts' design – data discretion for decomposing text-human image data (Sec. 4.2), and a new loss aiming to improve the alignment with respect to the scale of the human body and outfits (Sec. 8.3).

4.1 Preliminaries

We employ SD as the backbone model for its efficiency and widespread application in various downstream tasks. SD incorporates a variational autoencoder $\mathcal{E}$ to encode images $\mathbf{x}$ as latent variables z in a compact latent space, and applies diffusion schema in the latent space, thereby facilitating the diffusion process and reducing the computational cost. The denoising network is optimized by minimizing the L_2 error between predicted noise ϵ_θ and ground-truth noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathcal{L}_{noise} = E_{z\sim\mathcal{E}(x),c,\epsilon,t}\left[||\epsilon - \epsilon_\theta(z_t, t, c)||_2^2\right], \tag{1}$$

where z_t is the latent at time-step t, and c is the condition information that can be instantiated by text input.

The cross-attention layers are the hinge for textual information to play a role in influencing the updating of intermediate features. Specifically, a text prompt $\mathcal{P}$ is first transformed into a text embedding c via a CLIP text encoder. The latent z_t and text embedding are projected to form a query Q and keys K. The cross-attention maps are computed to flatten textual information into spatial features:

$$M = Softmax(\frac{QK^T}{\sqrt{d}}) \tag{2}$$

where d is the dimension of Q and K embeddings. The design works well when the text descriptions are short sparse captions. However, it can not handle text information with dense concepts, due to the lack of effective guidance to learn distinctive and precisely located features.

4.2 Data Discretization for Humans

We argue that there is no necessity to optimize the latent code guided by cross-attention maps during inference or harm the original architecture design of SD with sophisticated modules, *as long as keys K are decomposable and finite at the first place*. This is doable and simple to achieve. Because for the human generation scenario, textual descriptions about a person always revolve around body structure and attachments. As humans are structured in nature, textual descriptions could be explicitly classified into fixed groups that correspond to body regions, no matter how many concepts are described.

Thus, rather than directly utilizing nature language description, we propose a discretized textual prototype as illustrated in Fig. 6 for the network to enable precise communication between tokens. The prototype defines the convention of classifying and arranging the concepts of text captions to a finite set, where all captions can be represented as $C = \{C_{body}, C_{outfit}\}$. The subset C_{body} is for overall appearance, and the subset C_{outfit} is for fine-grained attributes of outfits.

Concretely, as shown in Fig. 6, given a human data sample $\mathbf{x}$ in CosmicMan-HQ, we first reorganize human parsing maps into the semantic map sets $H = \{h_i\}_{i=1}^N$, where N is the number of semantic masks and h_1 is the aggregation of all human parsing maps to differentiate human foreground with background. These masks are categorized into two levels – h_1 lies in *human-body-level* and the others belong to *outfit-level*. Then, we split the text captions with respect to H. Specifically, $C_{body} = c_{(s_1,e_1)}$ and $C_{outfit} = \{c_{(s_2,e_2)}, \ldots, c_{(s_N,e_N)}, c_{other}\}$. $c_{(s_n,e_n)}$ denotes the n^{th} sub-caption group related to the semantic map h_n, and s_n, e_n are the start and end indices of the concepts in the caption respectively. We gather the caption phrases without corresponding semantic masks as c_{other}, such as the caption for background. Note that, as our dataset naturally constructs annotation labels in a hierarchical manner, we can easily associate textual concepts with the semantic maps. For example, given a semantic map h_2 that represents the top clothing mask, we retrieve all the labels related to the top clothing and group them as a sub-caption c_2.

8.3 Decomposing and Refocusing Features

During training diffusion models, the denoising loss $\mathcal{L}_{noise}$ can ensure the content generative capability of the model, but it lacks explicit alignment constraints between the caption and image pixels, especially when encountering descriptions with dense concepts that cover very high information density. Thus, on the shoulders of discrete human data mentioned in Sec. 4.2, we propose a new loss – HOLA (short for **H**uman Body and **O**utfit Guided **L**oss for **A**lignment) to seamlessly decompose the cross-attention features in SD model and enforce attention refocusing without adding extra modules.

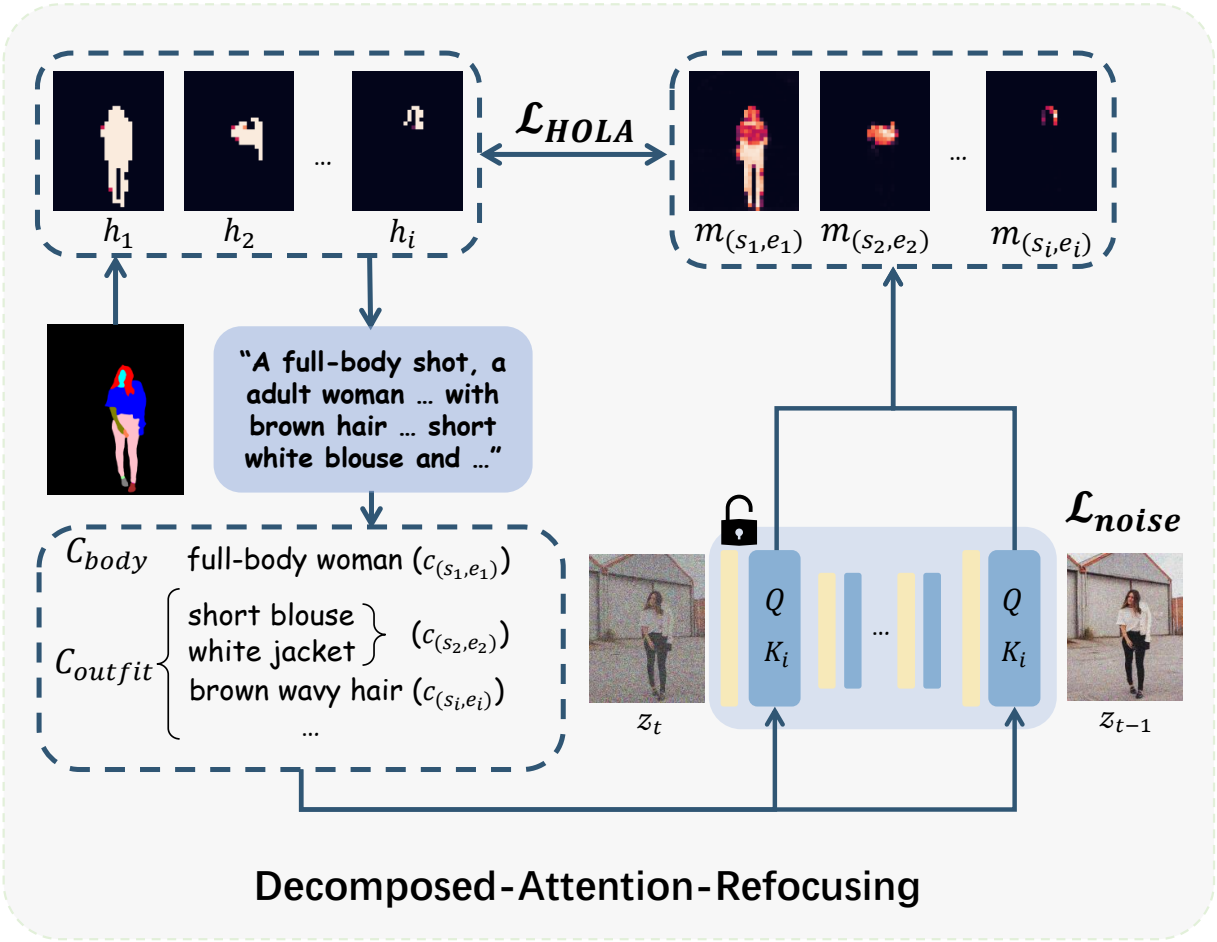

Figure 3: **Daring Training Framework.** It includes two parts: (1) data discretion for decomposing text-human data into fixed groups that obey human structure; (2) a new loss – HOLA, to enforce the cross-attention features actively response in proper spatial region with respect to the scale of body structure and outfit arrangement.

Figure 5: **Parsing Examples from CosmicMan-HQ.** The parsing results of sampled image in our dataset, along with detailed labels for each part. Text descriptions are obtained from labels.

Concretely, given the caption C and latent z_t, the cross-attention maps M can be decomposed as $M = (m_{(s_1,e_1)}, m_{(s_2,e_2)}, \ldots, m_{(s_N,e_N)}, m_{other})$. Each M_i is calculated through Eq. 2, with turning K to K_i (the projected embeddings of sub-caption $c_{(s_i,e_i)}$). We then incorporate HOLA alongside the original loss in SD to explicitly guide the cross-attention maps to have high responses only in specific regions. The HOLA is defined as follows:

$$\mathcal{L}_{\text{HOLA}} = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=s_i}^{e_i}\|m_j - h_i\|_2^2 + \left\|\frac{1}{e_i - s_i}\sum_{j=s_i}^{e_i}(m_j) - h_i\right\|_2^2) \tag{5}$$

Specifically, the first term of HOLA works under the guidance of human body structure – it pushes the high response region of each concept feature to be as close as possible to the corresponding semantic region. However, since certain outfit-related concepts may only occupy a specific proportion within a semantic region, it is unnecessary to enforce their features to align with the whole semantic region. Also, concepts within the same group should be arranged harmoniously. Thus, we use the second term of HOLA to satisfy the situation. This term requires the average attention maps within one group to be close to their semantic map. It helps reduce ambiguities in outfit-level descriptions. The overall loss function is as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{noise} + \beta\mathcal{L}_{\text{HOLA}} \tag{6}$$

where α and β are hyper-parameters to balance the contribution of each loss.

9 Experiments

In this section, we compare our method with SOTA text-to-image (T2I) methods on human-centric generation tasks. Sec. 9.1 shows experiment settings including implementation details and evaluation metrics. Sec. 9.2 validates our method outperforms state-of-the-art T2I methods from quantitative and human preference evaluation. Then, we provide an ablation study to evaluate the effectiveness of the design in training data and training strategy in Sec. 9.3. Finally, we show the pragmaticality and potential of CosmicMan as a foundation model in Sec. 9.4, by the application in two representative tasks in 2D and 3D respectively – 2D human image editing and 3D human reconstruction.

9.1 Experimental Settings

Implementation Details. Our foundation models are based on Stable Diffusion (SD-1.5 [43] and SDXL [37]). We finetune the whole UNet from the pretrained model of Stable Diffusion combined with Daring framework on CosmicMan-HQ 1.0 dataset. We use AdamW [32] as the optimized method in 1e-5 learning rate and 1e-2 weight decay. Our model is trained on 32 80G NVIDIA A100 GPUs in a batch size of 64 for about one week. Please refer to the supplementary material for more details.

Evaluation Metrics. We evaluate results from three perspectives: 1) Image Quality: Frechet Inception Distance (FID) [19] and Human Preference Score v2 (HPSv2) [54] are used to reflect diversity and authenticity. 2) Text-Image Alignment: CLIPScore [18] provides a holistic measure of image-text alignment. However, it struggles to capture detailed image-text relationships, especially in fine-grained texture, shape, and object descriptions, which also be discussed in [7, 21, 15]. Our proposed semantic accuracy metric, inspired by DSG [7], enhances fine-grained text-image alignment, focusing on object (Acc_{obj}), texture (Acc_{tex}), shape ($\text{Acc}_{\text{shape}}$), and overall ($\text{Acc}_{\text{all}}$), making it suitable for human-centric evaluation. 3) Human Preference: we conduct a user study to evaluate the image quality and text-image alignment of each method.

9.2 Comparison to Text-to-Image Models

We compared our foundation model with various state-of-the-art text-to-image models, including open-source models like Stable Diffusion (SD-1.5/2.0), SDXL, DeepFloyd-IF, and commercial models such as DALLE2/3 and MidJourney. For a thorough comparison, we evaluated two versions of our foundation model: CosmicMan-SD based on SD-1.5 and CosmicMan-SXDL based on SDXL.

Quantitative Evaluation. We prepared a test set comprising 2048 human images with fine-grained manually annotated prompts for fine-grained text-image generation. We report the quantitative comparison in Tab. 4. CosmicMan-SDXL excels in both image quality (FID) and fine-grained text-image alignment (Acc_{all}). In terms of image generation quality, CosmicMan-SD/SDXL outperforms the corresponding SD-1.5/SDXL by a large margin, showing up to 23.52% and 27.13% relative improvements in FID. As for fine-grained text-image alignment, CosmicMan-SDXL demonstrates superior generative capabilities in terms of three types of descriptions: object, texture, and shape. Compared to DALLE-3, which also unleashes the potential of detailed descriptions, CosmicMan shows 7.54%, 1.38%, 15.97% and 7.46% relative performance

Table 3: **Dataset Comparison.** The statistical comparison between publicly available human-related datasets and CosmicMan-HQ 1.0. "Common Scale" refers to the dataset that includes images captured at common scales, such as full-body shots, portrait photos, and half-body shots. "HP" and "Aes" refer to Human Parsing maps and Aesthetic scores.

	Data Quantity			Imaging Quality		Annotation							3*Domain
	Total Image #	Mean Resolution	Common Scale	Global↑	Face↑	Cat #	Attr #	Text	Bbox	Kpts	HP	Aes	
Human-Art [24]	$50K$	1115×1287	[HTML]A5B592	3.42	2.87	-	-	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	[HTML]D092A7	[HTML]D092A7	Real world & AI
DF-MM [22]	$44K$	750×1101	[HTML]D092A7	4.64	3.38	18	$587K$	[HTML]A5B592	[HTML]A5B592	[HTML]D092A7	[HTML]A5B592	[HTML]D092A7	Fashion
LAION-Human [25]	$1M$	688×650	[HTML]A5B592	4.20	2.66	-	-	[HTML]A5B592	[HTML]D092A7	[HTML]D092A7	[HTML]D092A7	[HTML]A5B592	Real world
SHHQ 1.0 [14]	$40K$	1024×512	[HTML]D092A7	4.23	2.13	-	-	[HTML]D092A7	[HTML]D092A7	[HTML]A5B592	[HTML]A5B592	[HTML]D092A7	Fashion
CosmicMan-HQ 1.0	$6M$	1488×1255	[HTML]A5B592	4.37	3.37	70	$115M$	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	[HTML]A5B592	Real world

boost on Acc_{obj}, Acc_{tex}, $\text{Acc}_{\text{shape}}$ and Acc_{all}. Note that CosmicMan-SD/SDXL obtains a relatively low CLIPScore, as our emphasis was on evaluating fine-grained text-image alignment. In contrast, CLIPScore lacks the ability for fine-grained evaluation, consistent with the conclusions in the DSG [7] and GenEval [15]. CosmicMan-SD/SDXL achieves the best performance in Acc_{all} and human preference evaluation, indicating its superiority in 2D human image generation.

Human Preference Evaluation. We compared our results with DeepFloyd-IF, SDXL, DALLE3, and MidJourney through pairwise comparisons. The evaluation considered both image quality and text-image alignment, using 100 randomly selected prompts to generate corresponding images for each method. The evaluation results exhibits a significant preference, with over 93.06%, 82.93%, 78.13% and 70.43% of subjects favoring our results in terms of image quality, and 85.38%, 90.25%, 88.56% and 81.68% of subjects preferring our results over those of DeepFloyd-IF, SDXL, DALLE-3, and MidJourney in terms of text-image alignment. Qualitative results in the supplementary material further highlight our model's superiority in image quality, fine-grained details, and text-image alignment.

9.3 Ablation Study

Ablation on Training Data. To show the validity of our proposed CosmicMan-HQ dataset, Tab. **??** reports the evaluation from three aspects: data source, data scale and annotation quality. 1) Data Source. Compared with two cutting-edge datasets, LAION-5B and HumanSD, *Ours* surpasses them by over 11.44 and 10.52 in FID, and 5.1 and 4.7 in Acc_{all}, respectively. LAION-5B has large noise in both data and annotation, while HumanSD has fewer data quantities and coarse annotations. Owing to the scalable ability of our data production workflow, Annotate Anyone, the constructed CosmicMan-HQ dataset features a large quantity of high-quality annotations, which benefits the final results. 2) Data Scaling. *Ours*, trained with $6M$ images, brings a promotion of 2.51 in FID and 0.9 Acc_{all} compared to $1M$ version *Ours-3*, proving the effectiveness of data scaling. Thus, Annotate Anyone's capacity to run constantly to produce data is necessary to push the boundaries of foundation models' performance. 3) Annotation Quality. We make a comparison under three different caption settings. *Ours-3* with AA caption exhibits a significant improvement of 7.57 and 10.94 in FID, as well as 3.6 and 3.2 in Acc_{all} compared to *Ours-1* and *Ours-2*. This verifies the effectiveness of improving the annotation quality of our proposed human-in-the-loop annotation mechanism in Annotate Anyone.

Ablation on Training Strategy. Tab. **??** shows the ablation of the training dataset and model design used in CosmicMan. By leveraging our CosmicMan-HQ dataset, fine-tuning the model gains a promotion of 10.52 in FID and 6.3 in Acc_{all}. Our proposed $\mathcal{L}_{\text{HOLA}}$ further enhances FID and Acc_{all} by 0.79 and 1.5. Our novel perspectives on data and model design boost remarkable promotions of CosmicMan on fine-grained human generation.

9.4 Applications

To validate the effectiveness of our human-specialized foundation model, we conduct additional experiments on 2D and 3D human-centric applications.

2D Human Editing. 2D human editing manipulates human images for specified poses. We compare our CosmicMan-SDXL with SDXL based on T2I-Adapter [34]. In Tab. **??**, our model outperforms SDXL on both FID and Acc_{all}, showing its superiority in 2D human editing tasks.

3D Human Reconstruction. We validate the effectiveness of our CosmicMan-SD model based on Magic123 [39], one representative 3D object reconstruction method from a single image. We replace the SD

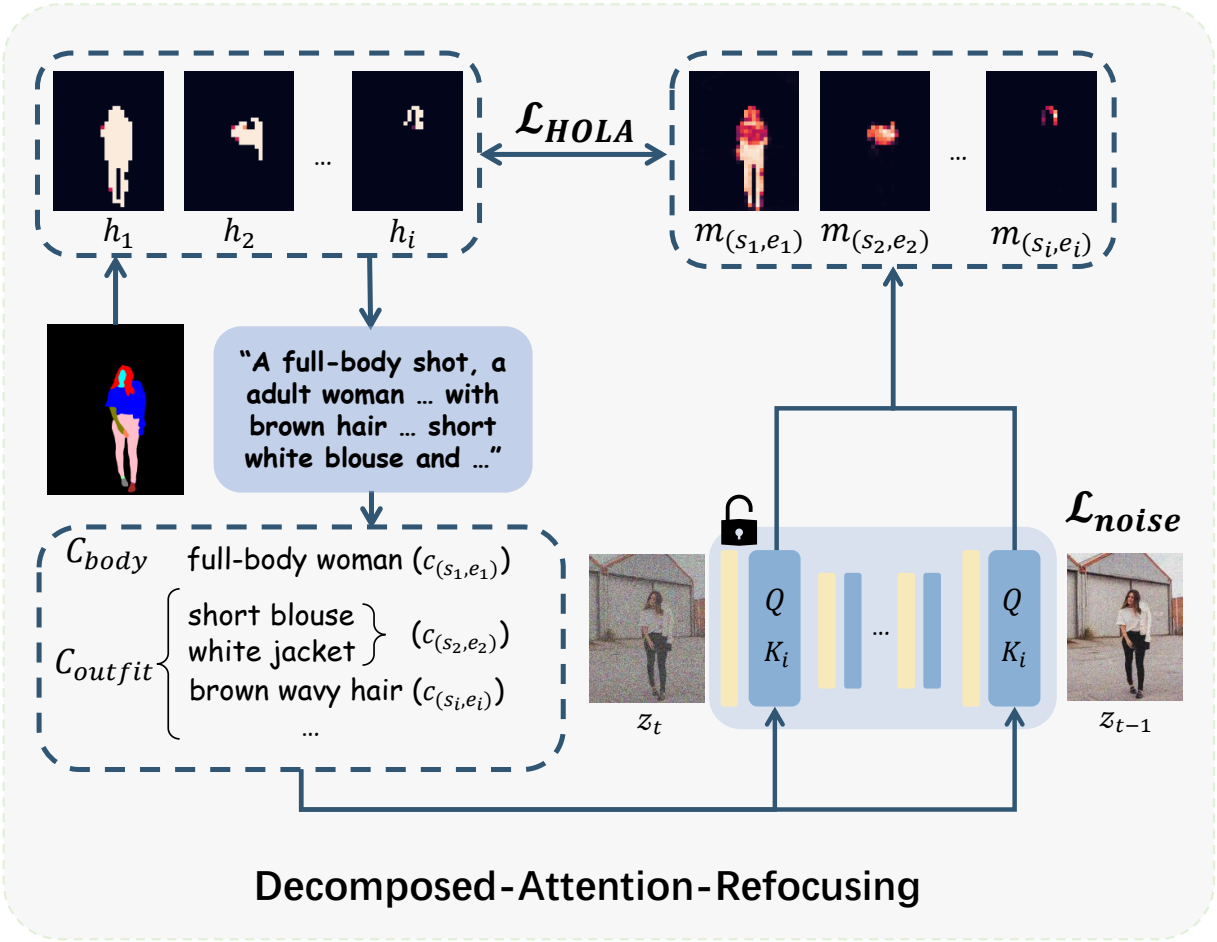

Figure 6: **Daring Training Framework.** It includes two parts: (1) data discretion for decomposing text-human data into fixed groups that obey human structure; (2) a new loss – HOLA, to enforce the cross-attention features actively response in proper spatial region with respect to the scale of body structure and outfit arrangement.

Table 4: **Quantitative Comparison to SOTA Text-to-Image Models.** The best and second-best results are marked with [HTML]E38AAERed and [HTML]85BB65Green.

1012c1012c1012c1012c1012c1012c1012c1012c

Methods	FID↓	HPSv2↑	CLIP↑	$\mathbf{Acc_{obj}}$↑	$\mathbf{Acc_{tex}}$↑	
SD 1.5 [43]	48.09	0.2659	[HTML]85BB6530.43	87.3	77.4	
SD 2.0 [43]	51.61	0.2588	26.27	82.8	74.7	
SDXL [37]	48.61	0.2647	[HTML]E38AAE30.78	88.5	82.5	
DeepFloyd-IF [9]	44.62	0.2603	29.33	87.9	84.4	
DALLE-2 [41]	49.60	0.2630	29.86	83.3	79.3	
DALLE-3 [2]	66.36	0.2673	28.86	86.2	87.1	
MidJourney [33]	53.89	0.2688	28.89	85.2	79.5	
CosmicMan-SD	[HTML]85BB6536.78	[HTML]85BB650.2690	28.47	[HTML]85BB6591.7	[HTML]85BB6585.7	[H
CosmicMan-SDXL	[HTML]E38AAE35.42	[HTML]E38AAE0.2698	27.31	[HTML]E38AAE92.7	[HTML]E38AAE88.3	[H

pretrained model with our foundation model in Magic123 for comparison. The higher CLIP-similarity [39] and $\mathrm{Acc_{all}}$ in Tab. **??** exhibit the superior potential of our model on 3D human reconstruction.

10 Discussion

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. The model and codes will also be released.

Future Work. Not placing CosmicMan merely as a research paper, we also commit ourselves to providing a long-term and sustainable foundation platform to support the research in human-centric content generation. Thus, we will continuously 1) operate Annotate Anyone to produce subsequent versions of CosmicMan-HQ aligned dynamically with real-world data, and 2) provide up-to-date human-specialized foundation models periodically trained on new versions of our data. By providing a well-constructed and long-term-maintained infrastructure, we hope to benefit broader research communities centered on human subjects.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. 2023. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[6] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19982–19993, October 2023.

[7] 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.

[8] 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.

[9] DeepFloyd. Deepfloyd-if, 2023. URL `https://github.com/deep-floyd/IF`.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[12] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] Flickr. Flickr application programming interface (api), 2023. URL `https://www.flickr.com/services/api/`. Accessed: 2023-11-18.

[14] 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.

[15] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023.

[16] Honglin He, Zhuoqian Yang, Shikai Li, Bo Dai, and Wayne Wu. Orthoplanes: A novel representation for better 3d-awareness of gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22996–23007, October 2023.

[17] 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.

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[21] 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.

[22] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. doi: 10.1145/3528223.3530104.

[23] Byungho 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.

[24] 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.

[25] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023.

[26] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.

[27] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide and bind your attention for improved generative semantic nursing, 2023.

[28] Anran Lin, Nanxuan Zhao, Shuliang Ning, Yuda Qiu, Baoyuan Wang, and Xiaoguang Han. Fashiontex: Controllable virtual try-on with text and texture. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[29] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV 2014*, 2014.

[30] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.

[31] 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.

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[33] Midjourney. Midjourney, 2023. URL `https://www.midjourney.com/`.

[34] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[35] Pixabay. Pixabay application programming interface (api), 2023. URL `https://pixabay.com/api/docs/`. Accessed: 2023-11-18.

[36] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.

[37] 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.

[38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[39] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, 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.

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*, pages 8821–8831, 2021.

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[42] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment, 2023.

[43] 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.

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, June 2023.

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[46] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019.

[47] 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.

[48] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. 2020.

[49] 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.

[50] Unsplash. Unsplash application programming interface (api), 2023. URL `https://unsplash.com/documentation`. Accessed: 2023-11-18.

[51] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2017.

[52] 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.

[53] Luozhou Wang, Guibao Shen, Yijun Li, and Ying cong Chen. Decompose and realign: Tackling condition misalignment in text-to-image diffusion models, 2023.

[54] 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.

[55] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

[56] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. 2023.

[57] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3dhumangan: 3d-aware human image generation with 3d pose mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23008–23019, October 2023.

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, October 2023.

[59] 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.

[60] Yutong Zhou and Nobutaka Shimada. Generative adversarial network for text-to-face synthesis and manipulation with pretrained bert model. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08, 2021.