

Human-Centric Foundation Models: Perception, Generation and Agentic Modeling

Shixiang Tang^{1,5}, Yizhou Wang¹, Lu Chen², Yuan Wang³,
Sida Peng², Dan Xu⁴, Wanli Ouyang^{1,5}

¹The Chinese University of Hong Kong, Hong Kong

²State Key Lab of CAD&CG, Zhejiang University, Zhejiang, China

³Tsinghua University, Beijing, China

⁴Hong Kong University of Science and Technology, Hong Kong

⁵Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract

Human understanding and generation are critical for modeling digital humans and humanoid embodiments. Recently, Human-centric Foundation Models (HcFMs)—inspired by the success of generalist models such as large language and vision models—have emerged to unify diverse human-centric tasks into a single framework, surpassing traditional task-specific approaches. In this survey, we present a comprehensive overview of HcFMs by proposing a taxonomy that categorizes current approaches into four groups: (1) Human-centric Perception Foundation Models that capture fine-grained features for multi-modal 2D and 3D understanding; (2) Human-centric AIGC Foundation Models that generate high-fidelity, diverse human-related content; (3) Unified Perception and Generation Models that integrate these capabilities to enhance both human understanding and synthesis; and (4) Human-centric Agentic Foundation Models that extend beyond perception and generation to learn human-like intelligence and interactive behaviors for humanoid embodied tasks. We review state-of-the-art techniques, discuss emerging challenges and future research directions. This survey aims to serve as a roadmap for researchers and practitioners working towards more robust, versatile, and intelligent digital human and embodiments modeling. Website is <https://github.com/HumanCentricModels/Awesome-Human-Centric-Foundation-Models/>

1 Introduction

Recent years have witnessed remarkable strides towards understanding human appearance, emotions, identities, actions, intentions and generating photorealistic humans in 2D and 3D. The success of these methods can be attributed to the robust estimation of identification [He *et al.*, 2024; Li *et al.*, 2024a], 2D keypoints [Wang *et al.*, 2023; Yuan *et al.*, 2024], fine-grained body-part segmentation [Tang *et al.*, 2023; Chen *et al.*, 2023a], depth [Khrodgar *et al.*, 2024], textual

descriptions [Chen *et al.*, 2024], and human meshes [Cai *et al.*, 2024], as well as the powerful human-centric deep learning frameworks, *e.g.*, vision transformers [Jin *et al.*, 2024; Wang *et al.*, 2023; Huang *et al.*, 2024a] and diffusion models [Ju *et al.*, 2023; Li *et al.*, 2024b; Lin *et al.*, 2025]. Despite progress in every individual task, robust and accurate understanding and generating photorealistic and even intelligent digital humans requires to deeply understand human as a holistic and complex system at the intersection of diverse human-centric tasks related to appearance, identity, motion and intentions. Moreover, most existing human-centric pipelines are task-specific for better performances, leading to huge costs in representation/network design, pretraining, parameter-tuning, and annotations. Therefore, the recent human-centric learning community appeals for a unified framework [Ci *et al.*, 2023; Wang *et al.*, 2023; Chen *et al.*, 2024; Huang *et al.*, 2024a] to unlock systematic understanding and a wide range of human-centric applications for everybody.

Inspired by rapid advancements of general foundation models, *e.g.*, large language models (LLMs), large vision models (LVMs) and text-to-image generative models, and their presents of a paradigm shift from end-to-end learning of task-specific models to generalist models, a recent trend is to develop **Human-centric Foundation Models (HcFM)** that satisfy three criteria, namely generalization, broad applicability, and high fidelity. Generalization ensures robustness to unseen conditions, enabling the model to perform consistently across varied environments. Broad applicability indicates the versatility of the model, making it suitable for a wide range of tasks with minimal modifications, or even without modifications. High fidelity denotes the ability of the model to produce precise, high-resolution outputs essential for faithful human generation tasks such as 2D to 3D lifting. Recent notable works towards such human-centric foundation models include SOLIDER [Chen *et al.*, 2023a], PATH [Tang *et al.*, 2023], UniHCP [Ci *et al.*, 2023], Sapiens [Khrodgar *et al.*, 2024], MotionGPT [Jiang *et al.*, 2023; Zhang *et al.*, 2024a], ChatHuman [Lin *et al.*, 2024], *etc.*

In light of the rapid developments and emerging challenges of Human-centric foundation models, we present a comprehensive survey of this field to help the community keep track of its progress. Specifically, we introduce a taxonomy that categorizes existing works into four groups according to their sup-

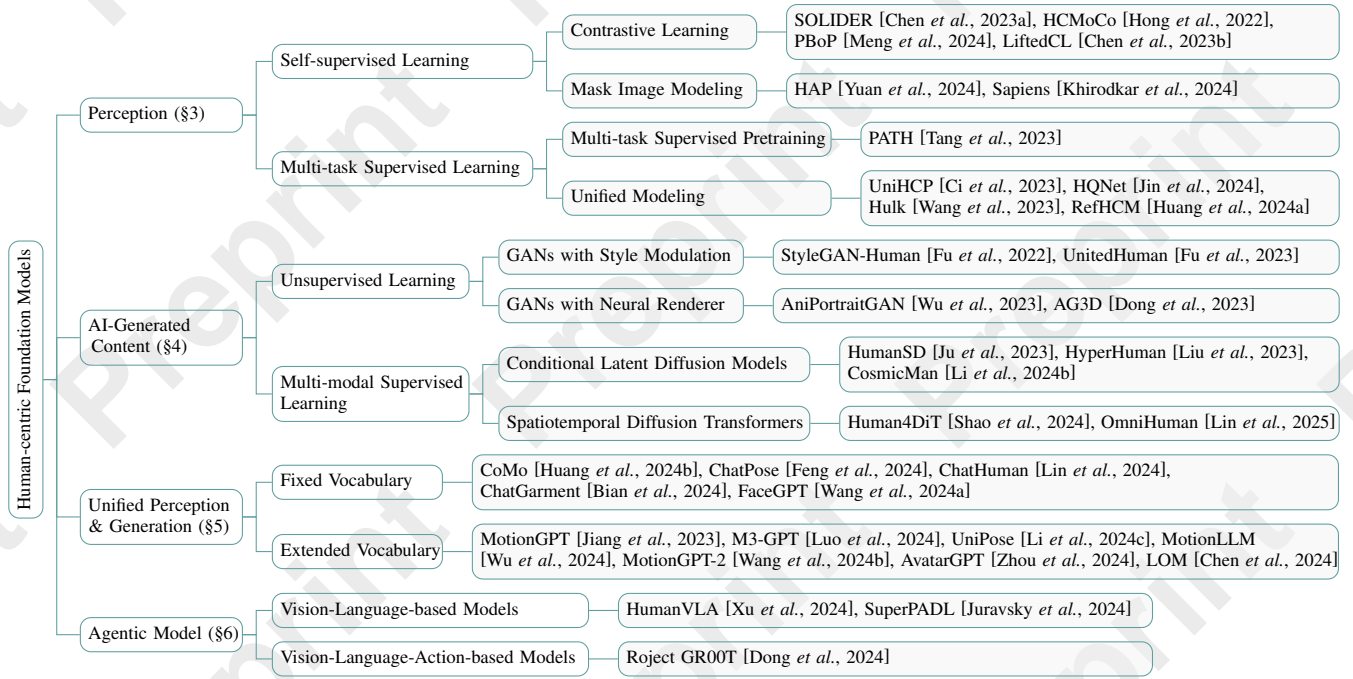


Figure 1: A taxonomy of human-centric foundation models with representative examples.

ported downstream tasks: *Human-centric perception foundation models*, *Human-centric AIGC foundation models*, *Human-centric unified perception and generation foundation models*, and *Human-centric agentic foundation models*. Human-centric perception foundation models modify the existing unsupervised and multitask supervised pretraining framework to capture the fine-grained human-centric features essential to multi-modal 2D and 3D perception tasks, such as skeleton-based action recognition, human parsing, *etc.* Human-centric AIGC foundation models build on the success of generative foundation models by leveraging rich human-specific data, focusing on creating human-related content with high realism and diversity. To move a step forward, human-centric perception and generation tasks can be unitedly modeled in the single foundation model with the inspiration of multi-modal large language models. Such unified modeling are proven not only to benefit human understanding but also fine-grained and photorealistic generation. Lastly, as mentioned by Michael J Black and Xavier Puig [Feng and others, 2024], the human foundation agent model goes beyond perception and generation, it is a model that can receive signals from human sensors, and interact with humans based on the inputs and, for which, the embodiment, actions, and motor behaviors are human-like. Different from other foundation models, the human-centric agent foundation models aim at learning human intelligence, which hopefully can benefit various embodied AI tasks.

To our best knowledge, this is the first survey about human-centric foundation models with a novel taxonomy (Fig. 1). Its scope also extends beyond the mere categorization of existing techniques. It also explores the potential future trajectories, contemplating how ongoing advancements might unfold, including data, ethics and technological aspects.

2 Taxonomy

The objective of the taxonomy is to group Human-centric foundation models with the similar supported tasks into the same category. Specifically, we classify existing human-centric foundation models into four categories, *i.e.*, Human-centric perception foundation models, Human-centric AIGC foundation models, Human-centric unified perception and generation foundation models, and Human-centric agentic models, which can be further summarized in difference learning frameworks. We briefly introduce the four categories as follows:

(1) **Human-centric perception foundation models** learn from large-scale and multi-modal human-centric data and support major perceptive tasks, including person re-identification, human parsing, pose estimation, human mesh recovery and skeleton-based action recognition. The fundamental idea to leverage structures of human bodies or diverse annotations to learn the fine-grained and semantic human-centric representations. Based on the learning frameworks, human-centric perception foundation models can be further categorized into unsupervised learning and multitask supervised learning.

(2) **Human-centric AIGC foundation models** are designed to create content—such as images, videos, or avatars—that centers on human. These models are trained on extensive human-centric data to produce realistic and diverse portrayals of individuals, whose objective is to produce content that accurately reflects fine-grained human appearances, behaviors, and interactions. Based on the training paradigm, these models can be broadly classified into models trained with unsupervised learning and multi-modal supervised learning.

(3) **Human-centric unified perception and generation foundation models** can support perception and generation tasks closely correlated. The fundamental idea is to view human-centric cues other than texts as foreign languages and append

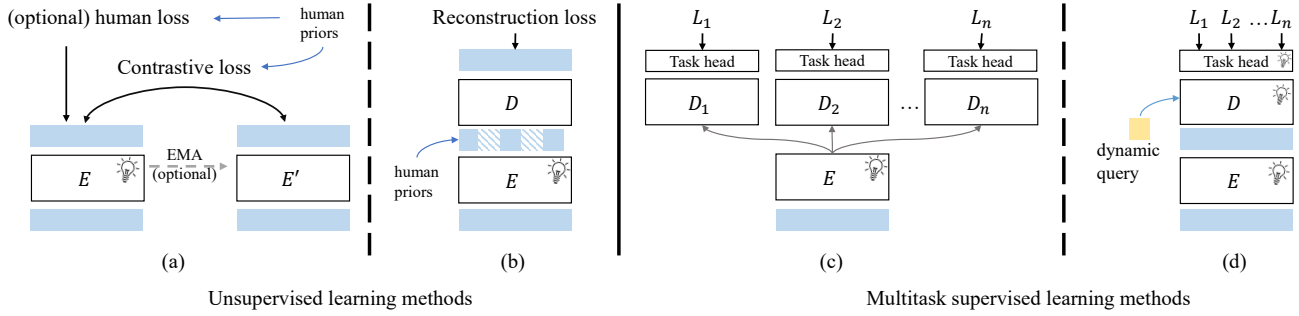


Figure 2: Different Frameworks of human-centric perception foundation models. Parameters in modules with ⚡ are used in downstream tasks. Unsupervised foundation models: (a) Contrastive learning methods; (b) Mask image modeling methods. Supervised foundation models: (c) Multitask supervised pretraining methods; (d) Unified modeling methods.

them to large language models (LLMs) as multi-modal large language models (MLLM) for understanding and generation. **(4) Human-centric agentic foundation models** learn human intelligence and support human-centric embodied AI tasks, such as human-robot collaboration tasks and social interactions. Based on the learning frameworks, human-centric agentic foundation models can be categorized into vision-language-based method and vision-language-action-based methods.

The four categories of foundation models are interconnected rather than mutually exclusive, contributing collectively to the rapid advancement of the human-centric foundation models. In the follows, we delve into each of them, exploring key challenges, representative solutions, and emerging trends.

3 Human-centric Perception Foundation Models

The human-centric perception foundation model demonstrates that compact ahuman-centric representations can be learned from multiple human-centric task and efficiently adapted to a broad range of 2D and 3D multi-modal perception tasks. Based on whether to leverage human-centric annotations, we cover two paradigms of human-centric perception foundation models: unsupervised learning and supervised learning.

3.1 Unsupervised learning methods

Human-centric unsupervised foundation models were proposed to mitigate dependency on the sensitive annotations, which mainly follow the pretraining-finetuning paradigm. During pretraining, without labels, they used the inherent priors of human body structure to learn versatile and representative human-centric features in the encoder. When adapted to downstream tasks, the pretrained encoder and the task head will be parameter-efficiently and fully finetuned using labeled data.

Contrastive learning methods leverage human priors to align features derived from encoders using tailored contrastive losses, as shown in Fig. 2(a). Considering the multi-modal nature of human data (e.g., RGB, depth, 2D keypoints), instead of commonly used momentum encoders, multiple encoders were used to process inputs with different modalities. Human priors were included to guide multi-modal contrastive learning for general human-centric representations. HCMoCo [Hong *et al.*, 2022] employed multiple encoders to exploit multi-modal human body consistency through a hierarchical contrastive learning framework. Based on it, PBoP [Meng *et al.*,

2024] introduced an additional encoder for generated latent part pair images. The extracted features could then serve as anchors to guide the multi-modal contrasting learning process. However, the acquisition of multi-modal data is not always straightforward. When only images were available, integrating an additional loss with human prior knowledge was another effective solution. For example, SOLIDER [Chen *et al.*, 2023a] proposed an additional semantic classification loss to import semantic information into the learned features. LiftedCL [Chen *et al.*, 2023b] introduced an adversarial loss to supervise the lifted 3D skeletons, explicitly inserting 3D human structure information for human-centric pretraining.

Mask image modeling methods implicitly learn human knowledge by reconstructing masked inputs (see Fig. 2(b)) based on the prior knowledge of body structures. HAP [Yuan *et al.*, 2024] utilized the 2D keypoints to guide the mask sampling process during mask image modeling, encouraging the model to concentrate on body structure information. To introduce 3D human prior, [Armando *et al.*, 2024] proposed to reconstruct masked pedestrian images from cross-view and cross-pose pairs. Thanks to human priors, human-centric unsupervised foundation models show superior performance than the ImageNet pretraining methods on human-centric perception tasks, especially on low-data regimes.

3.2 Multitask supervised learning

When labeled data is sufficient, multitask supervised learning can exploit the internal relationship among data, emerging as a straightforward and effective paradigm for constructing human-centric perception foundation models. Some recent works learned the shared information among different human-centric datasets to benefit specific human-centric tasks, e.g., human shape estimation [Cai *et al.*, 2024], pedestrian detection [Zhang *et al.*, 2024b], re-identification [He *et al.*, 2024; Li *et al.*, 2024a]. By constructing a unified framework applicable to related subtasks and training them concurrently, these approaches have outperformed specialist models. However, a notable limitation of these methods is their inability to learn and execute other perception tasks, which restricts the potential scope of the application. In response to the challenge, recent advances in multiple human-centric tasks co-training showed a new direction. These developments demonstrate that human-centric perception foundation models can exploit inter-task homogeneity to enhance overall performance.

Multitask supervised pretraining methods centered on utilizing multiple distinct supervisions to force the encoder to learn general human-centric representation (see Fig. 2(c)). To handle task conflicts brought by supervised learning from diverse labels, PATH [Tang *et al.*, 2023] leveraged task-specific projectors along with the hierarchical weight-sharing strategy, enforcing the encoder to acquire general representations for downstream human-centric tasks.

Unified modeling methods have emerged as the mainstream in human-centric perception foundation models to further mitigate the resource-intensive full finetuning process for downstream tasks. As depicted in Fig. 2(d), these methods followed the unified encoder-decoder framework with dynamic queries. UniHCP [Ci *et al.*, 2023], as a first attempt, adopted a task-guided interpreter to unify task heads and task-specific queries, handling five human-centric tasks. HQNet [Jin *et al.*, 2024] focused on instance-level features for individual persons and proposed Human Queries to learn unified all-in-one query representations, providing a single-stage method to tackle multiple distinctive human-centric tasks. While these works mainly concentrate on 2D human-centric tasks, Hulk [Wang *et al.*, 2023] extended the scope to simultaneously address 2D vision, 3D vision, vision-language and skeleton-based human-centric tasks. To tackle these tasks, Hulk categorized the input and output formats into four modalities and developed modality-specific (de-)tokenizers with modality indicator queries, unifying all tasks into modality translation tasks. Considering human-computer interaction, RefHCM [Huang *et al.*, 2024a] converted multi-modal data into semantic tokens, unifying various perception tasks as referring tasks.

4 Human-centric AIGC Foundation Models

Human-centric AIGC foundation models are generative models specifically trained or fine-tuned on *large-scale human datasets* to create content focused on human elements, such as images, videos, and 3D avatars. By prioritizing the generation fidelity of human attributes, these models serve as *adaptable bases for downstream tasks* like image editing, virtual try-on, 3D generation, and character animation. In this section, we categorize these models based on unsupervised learning and multi-modal supervised learning approaches.

4.1 Unsupervised Learning

Unsupervised learning plays a crucial role in early human-centric generative models, with GAN-based methods leading the progress in realistic human image and avatar generation. Specifically, there are two main frameworks: *2D GANs with Style Modulation* and *3D-Aware GANs with Neural Renderer*. **GANs with Style Modulation** leverage an intermediate, disentangled latent space to exert fine-grained control over the image synthesis process by first mapping a random noise to a style vector and then using learned affine transformations to inject style vectors at various layers of the generator, as shown in Fig.3 (a). This hierarchical modulation allows fine-grained control of visual attributes at different scales. For instance, StyleGAN-Human [Fu *et al.*, 2022] trained a series of unconditional models with this framework and showed that scaling data size, balancing data distribution, and aligning human body could significantly enhance generation performance, thereby establishing robust foundational models for

conditional applications. UnitedHuman [Fu *et al.*, 2023] employed a multi-source spatial transformer to align multi-source data, including face, hand, partial-body, and whole-body images, into a unified space for more comprehensive human modeling, and designed a continuous generator to synthesize coherent high-resolution images with enhanced details. These models served as versatile, pre-trained foundations for a variety of downstream applications. Their well-structured latent space enabled (1) intuitive drag manipulation of image attributes [Pan *et al.*, 2023] by optimizing latent codes based on point movement supervision, (2) 3D generation [Xiong *et al.*, 2023] with reconstruction priors to produce disentangled geometry and texture code, and (3) virtual try-on [Yoshikawa *et al.*, 2023] by aligning the latent code with garment features.

GANs with Neural Renderer integrate 3D representations and a neural renderer into the former framework to enforce 3D geometric consistency during image synthesis, (Fig.3(b)). By guiding the generator with rendering results (e.g., low-resolution image, depth, or surface normals), these models produce outputs that better reflect realistic spatial structures, making them especially to create high-fidelity 3D avatars. For example, SofGAN [Chen *et al.*, 2022] decoupled the 3D representation space into geometry and texture subspaces, providing a robust foundation for independent control over camera pose, facial structure, and attribute texture in various human portraits. AniPortraitGAN [Wu *et al.*, 2023] integrated human priors into its framework by learning pose and facial deformations using the SMPL model and 3D Morphable Model (3DMM). Trained on large-scale facial image collections, its generator and renderer could serve as the basis for tasks that require high-resolution controllable images with detailed 3D geometry. Similarly, AG3D [Dong *et al.*, 2023] extended this framework to train on large-scale full-body images by incorporating an additional deformer for learning pose-dependent effects and a normal estimator for geometry supervision, resulting in high-quality 3D human avatars. These models' 3D-aware generation ability unlocked many downstream tasks, including (1) free-viewpoint video synthesis by rendering from arbitrary camera trajectories, (2) pose retargeting by integrating learnable pose-dependent deformations, and (3) single-view 3D reconstruction with GAN inversion.

4.2 Multi-modal Supervised Learning

Recent advances in diffusion models have spurred the development of multi-modal supervised learning approaches that leverage large-scale paired datasets, such as text-image or video-pose alignments, to achieve precise control over the generation process. With *Conditional Latent Diffusion Models* and *Spatial-Temporal Diffusion Transformer Architecture*, recent works have significantly improved the quality, consistency, and diversity of synthetic human images and videos.

Conditional Latent Diffusion Models extend standard diffusion models by operating in a compact latent space and incorporating external conditions, such as text, pose, or segmentation maps, to guide the generation process (Fig.3 (c)). By first encoding inputs into a lower-dimensional latent representation, diffusion is applied to gradually denoise the output while preserving the provided conditions. This approach enhances both efficiency and controllability, making it a general framework for tasks like human image synthesis, pose-guided

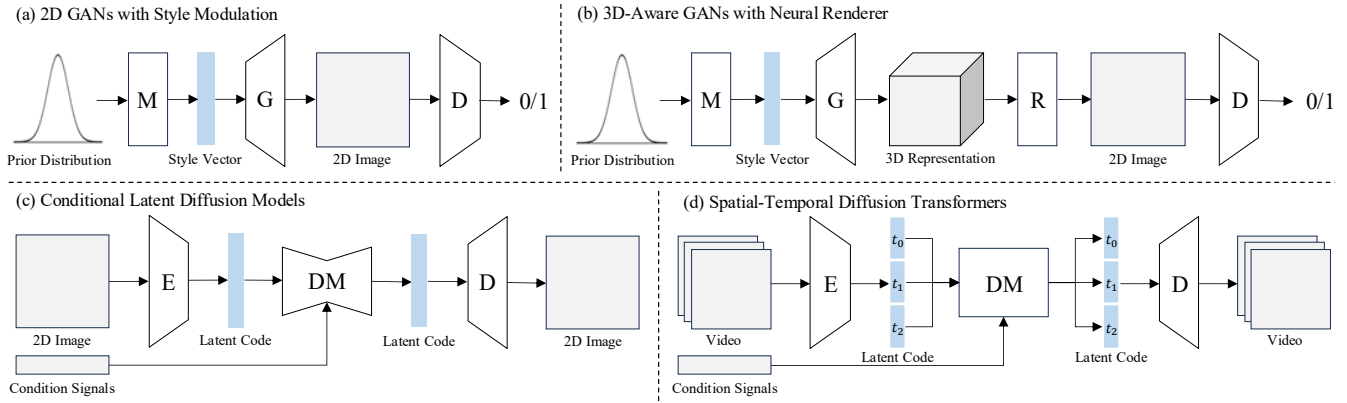


Figure 3: Different frameworks of human-centric AIGC foundation models. Unsupervised learning methods: (a) A sampled noise is transformed by a mapping network (M) into a style code that modulates the generator (G) at multiple scales, while a discriminator (D) is trained simultaneously through adversarial learning to refine realism. (b) 3D representations are incorporated along with a neural renderer (R) to enhance spatial consistency. Supervised learning methods: (c) The input image is encoded (E) into a latent code, which the diffusion model (DM) iteratively denoises to reconstruct high-quality outputs with fine-grained control. (d) The input video is compressed and decomposed into space-time latents, where transformers capture spatial-temporal dependencies and enable efficient scaling for video generation.

generation, and multi-modal content creation. Specifically, HumanSD [Ju *et al.*, 2023] proposed a skeleton-guided diffusion model with a heatmap-guided denoising loss. Trained on 2M+ text-image-pose triplets, this model demonstrated its foundational ability in generating high-quality human images across various scenarios. HyperHuman [Liu *et al.*, 2023] presented a unified framework for generating hyper-realistic human images by capturing correlations between appearance and structure in multi-modal data. Specifically, it introduced a latent structural diffusion model that jointly denoises depth, surface normal, and RGB images conditioning on caption and pose skeleton, where modality-specific branches can be complementary to each other. Recently, CosmicMan [Li *et al.*, 2024b] was proposed by building on three key pillars: a scalable, high-quality data production paradigm, robust model design via a decomposed attention-refocusing framework, and pragmatic integration into downstream tasks. This structure empowered CosmicMan to generate images in various scenarios, from full-body portraits to close-up shots, establishing it as a versatile cornerstone for human-centric content generation.

Spatial-Temporal Diffusion Transformers extend diffusion models by incorporating transformer-based architectures to capture both spatial structures and temporal dependencies in videos. By leveraging self-attention mechanisms across spatial dimensions and time steps, this framework (Fig.3 (d)) effectively captures long-range temporal relationships while maintaining spatial consistency in human-centric video generation. Consequently, it is ideally suited for tasks such as character animation, video manipulation, and 4D generation. A representative work is Human4DiT [Shao *et al.*, 2024], which introduced a hierarchical 4D Diffusion Transformer (DiT) for generating high-quality, 360-degree spatial-temporally coherent human videos from a single image. Trained on a multi-sourced dataset spanning images, videos, multi-view captures, and 4D footages, the model factorized self-attention across views, time, and space while incorporating accurate condition injection, successfully handling complex motions and viewpoint changes. Building on this framework, OmniHuman [Lin *et al.*, 2025] proposed a scalable, multi-modality-conditioned human

video generation model from a single image and motion signals (e.g., audio, video, or both). By leveraging mixed data and incorporating motion-related conditions during training, it mitigated data scarcity and enabled realistic video synthesis across diverse scenarios, including talking, singing, varying body compositions, and human-object interactions. Notably, pre-trained human-centric diffusion models are increasingly serving as the versatile foundation for a wide range of applications. For example, they can be (1) fine-tuned for text-driven image editing [Brooks *et al.*, 2023] using instruction-image pairs, (2) integrated into character animation [Hu, 2024] by incorporating an additional branch to merge detailed textures into character’s movements, (3) adapted for virtual try-on [Karras *et al.*, 2024] by splitting classifier-free guidance into person, garment, and pose conditioning, and (4) extended to 3D/4D human generation [Kolotouros *et al.*, 2023] by iteratively applying score distillation sampling in multiple views.

5 Human-centric Unified Perception and Generation Foundation Models

In recent years, human-centric foundation models emerge as a transformative approach for unifying perception and generation tasks, providing comprehensive frameworks to understand and synthesize human behaviors, motions, emotions, and intentions. By integrating multi-modal human-centric cues—such as visual, auditory, textual, emotional, and motion data—into LLMs and MLLMs, these models facilitate richer, context-aware representations of human interactions. Such models can be categorized into two primary paradigms—*fixed vocabulary* and *extended vocabulary*—based on how multi-modal human-centric signals are integrated into LLMs and MLLMs.

5.1 Fixed Vocabulary

Vocabulary-Fixed models enhance LLMs by introducing modality-specific projection layers to map human-centric signals into the feature space of LLMs or directly employing off-the-shelf tools and prompt engineering technique. Among these approaches, CoMo [Huang *et al.*, 2024b] unified the

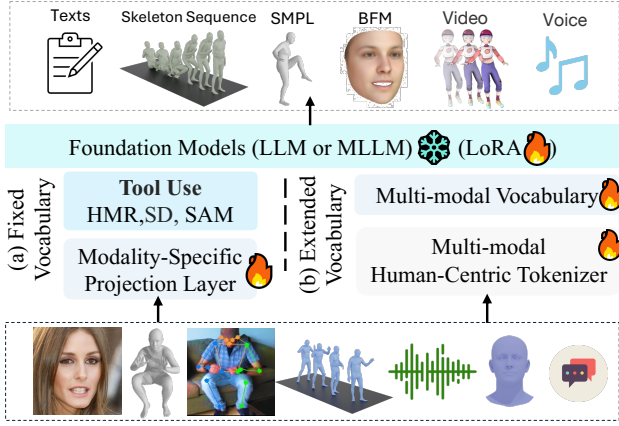


Figure 4: Human-centric foundation models for unified perception and generation tasks. (1) Vocabulary-Fixed models with fixed vocabulary introduces modality-specific projection layers or directly employing off-the-shelf human-centric tools. (2) Vocabulary-extended models align enriched human-centric representations, e.g., pose parameters, SMPL representations, motion sequences, emotions, and audio signals with the original vocabulary of foundational LLMs.

text conditioned human motion generation, fine-grained motion generation and motion editing. Specifically, it autoregressively generates sequences of interpretable pose codes upon the high-level text description and fine-grained, body-part-specific descriptions generated by LLMs. ChatPose [Feng *et al.*, 2024] introduced LLMs to advance pose-related tasks, striving to develop a versatile pose generator. By integrating image interpretation, world knowledge, and body language comprehension into foundational LLMs, ChatPose enhanced its ability to understand and reason about 3D human poses from images and textual descriptions. Taking this a step further, ChatHuman [Lin *et al.*, 2024] introduced a multi-modal LLM integrated with 22 domain-specific human-centric tools, improving its ability to reason about human-related tasks. Assisted by academic publications and retrieval-augmented generation model, ChatHuman generated in-context-learning examples for handling newly introduced tools. In other human-centric tasks, foundational models have also made remarkable strides. For example, ChatGarment [Bian *et al.*, 2024] leveraged large vision-language models (VLMs) to automate the estimation, synthesis, and editing of 3D garments from either images or textual descriptions. Similarly, FaceGPT [Wang *et al.*, 2024a] integrated 3D morphable face model (3DMM) [Blanz and Vetter, 2023] parameters into token space of VLMs, enabling the self-supervised generation of 3D facial models from both textual and visual inputs.

5.2 Extended Vocabulary

Vocabulary-extended models are designed to expand capabilities of LLMs [Touvron *et al.*, 2023; Achiam *et al.*, 2023] and MLLMs [Liu *et al.*, 2024; Li *et al.*, 2023] by explicitly extending their vocabulary and embedding spaces to accommodate human-centric and multi-modal signals. By aligning enriched human-centric representations—such as *pose parameters*, *SMPL representations*, or *motion sequences*—with the original vocabulary of LLMs, Vocabulary-Extension Mod-

els empower foundational LLMs and MLLMs to effectively tackle novel human-centric tasks, including generating, editing, and comprehending human poses of textual descriptions, image, and 3D modalities. Recent studies [Wu *et al.*, 2024; Wang *et al.*, 2024b; Jiang *et al.*, 2023; Luo *et al.*, 2024; Zhou *et al.*, 2024] incorporated motion modality into textual space of the foundational LLMs, enabling a holistic representation of the complex relationship between motion and natural language. MotionGPT [Jiang *et al.*, 2023] introduced a unified motion-language model to handle multiple motion-related tasks. It first discretized continuous motions into discrete semantic tokens, which could be interpreted as “*body language*” and then be extended to the LLM’s vocabulary. Through pre-training alignment and prompt tuning stages, MotionGPT showcased strong generalization across various human motion understanding and generation tasks. Building on the MotionGPT [Jiang *et al.*, 2023], AvatarGPT [Zhou *et al.*, 2024] proposed an *all-in-one* structure for motion understanding, planning, generation, as well as motion in-between synthesis. M3-GPT [Luo *et al.*, 2024] further embed motion, music, and language into a single vocabulary and includes music-to-dance and dance-to-music tasks. Such a unified approach advanced human-centric tasks by synthesizing multi-modal human behaviors and bridging the gap between auditory, visual, and linguistic modalities. While these methods mostly were restricted to single-human motion, MotionLLM [Wu *et al.*, 2024] introduced a simple yet versatile framework capable of handling single-human, multi-human motion generation, and motion captioning by fine-tuning pre-trained LLMs. MotionGPT-2 [Wang *et al.*, 2024b] developed a general-purpose Large Motion-Language Model (LMLM), which extended beyond current solutions by tackling the challenging 3D holistic motion generation on the MotionX [Lin *et al.*, 2023] benchmark. However, these methods struggled with pose-related editing and comprehension. UniPose [Li *et al.*, 2024c] leveraged generation abilities of LLMs to unify all pose-relevant tasks to comprehend, generate, and edit human poses across various modalities, including images, text, and 3D SMPL representations. As a solution, LOM [Chen *et al.*, 2024] unified verbal and non-verbal language using MLLMs for human motion understanding and generation, accommodating text, speech, motion, or any combination thereof as input flexibly. It trained a compositional body motion VQ-VAE to tokenize motions into part-wise discrete tokens, unifying modality-specific vocabularies (audio and text).

6 Human-centric Agentic Foundation Models

Beyond understanding and generating human themselves, Human-centric agentic foundation models often process multi-modal inputs (e.g., egocentric images, tactiles, sounds, and natural language as task description) focus on the human reaction and interactions to the environments while being constrained by its physical properties. Leveraging the Internet-scale multi-modal dataset, Human-centric agentic foundation models hold the promise of generalization across diverse tasks and provide a natural interface for human-robot interaction, both essential for real-world robot applications.

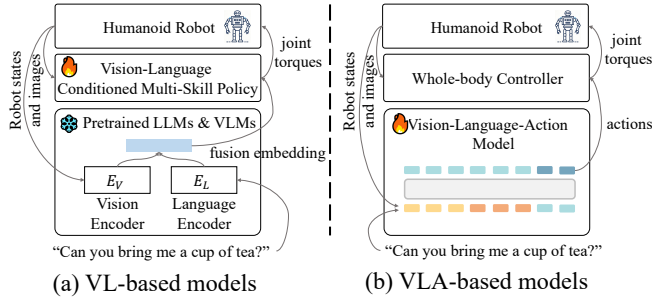


Figure 5: Frameworks of human-centric agentic foundation models to endow humanoid robots. (a) Vision-Language-based models: Use pretrained vision-language to learn multi-skill policy; (b) Vision-Language-Action-based models: Finetune a vision-language-action model to generate actions.

6.1 Vision-language-based models

Vision-language-based Models are prominent approaches in constructing foundation models for humanoid systems by integrating pre-trained vision-language models (VLMs) with sensorimotor control policies. As illustrated in Fig. 5(a), these methods leverage the strengths of visual and linguistic modalities to bridge high-level semantic understanding with low-level physical actions. In these frameworks, visual inputs are processed using pre-trained image encoders while natural language commands are tokenized through pre-trained text encoders. The extracted semantic features are then aligned with humanoid-specific control policies, enabling the translation of language and visual cues into precise motor actions. This cross-modal alignment not only facilitates language-to-action mapping but also enhances the capability of humanoid robots to perform complex, multi-step tasks. Recent efforts have focused on adapting these vision-language-based models to the unique challenges of humanoid robotics. Notable examples include HumanVLA [Xu *et al.*, 2024] and SuperPADL [Juravsky *et al.*, 2024], which aligned humanoid control policies with the latent spaces of pre-trained VLMs. These approaches have demonstrated the potential to endow humanoid robots with sophisticated skills that are driven by image and language inputs, paving the way for more advanced and natural human-robot interactions in dynamic environments.

6.2 Vision-language-action-based models

Vision-language-action (VLA) models have emerged as a promising approach in constructing foundation models for humanoid robots by unifying visual, linguistic, and action modalities within a single framework. As depicted in Fig. 5(b), these methods treat robot data—both observations and actions—as tokens in the vocabulary of a pre-trained language model, thereby enabling direct fine-tuning or co-training with established vision-language models. In contrast to traditional methods that depend on separately trainable low-level control policies, VLA models directly generate actions as token sequences, effectively merging high-level semantic reasoning with motor command generation. This unified tokenization framework allows the model to generate a wide range of skills while preserving robust language and vision understanding, thus enhancing its generalization across diverse tasks. How-

ever, representing actions as stringified tokens may become inefficient for high degrees-of-freedom systems such as humanoid robots, where the complexity of motor commands is significantly higher. Despite the potential benefits, applying the VLA paradigm to humanoid robotics remains largely unexplored. To address this gap, recent initiatives like NVIDIA’s Project GR00T [Dong *et al.*, 2024] have been announced. The GR00T foundation model aspired to leverage a diverse array of data sources—from internet and simulation data to real-robot interactions—to facilitate scalable training and achieve robust, cross-modal performance for complex humanoid tasks.

7 Challenges and Future Directions

Data. Different from general images and videos, collecting high-quality human-centric data is much more sensitive, difficult and expensive, compared to general images and videos. This situation inevitably leads to a trade-off in data quantity and data quality. Furthermore, the wide variability in human appearances, behaviors and contexts makes it hard to get a comprehensive dataset, limiting the generalization ability of data-driven human-centric foundation models.

Representations. Human appearance and behavior demand a holistic understanding that integrates body, face, and hands, yet no existing foundation model captures these aspects simultaneously. Recognizing that these elements are closely interrelated parts of a complex system, there remains a significant gap in unified modeling frameworks. Future research should focus on developing innovative, multi-modal architectures and scalable datasets that bridge the global and granular representations of human dynamics, ultimately enhancing applications in digital human synthesis, interactive robotics, and personalized human-computer interaction.

Interactivity. Despite significant progress in understanding and generating isolated human attributes such as appearance, emotion, and identity, current human-centric foundation models face challenges in capturing the complex interplay of interactions and contextual dynamics inherent in real-world scenarios. Most approaches are limited to static or isolated representations, hindering their ability to model the nuanced interdependencies between individuals and their environments. Future research should focus on developing unified frameworks that seamlessly integrate high-level semantic reasoning with fine-grained behavioral synthesis, enabling models to adapt to dynamic, multi-agent contexts.

Ethics. Ethics are crucial in applying human-centric foundation models, especially regarding sensitive domains and privacy. In sensitive domains, the model’s output should never replace the expertise of human professionals. Moreover, Human-centric foundation models should also be developed with privacy-enhancing technologies, ensuring that the model learns powerful representations from the data while making it difficult to identify individual-level privacy information. Additionally, anonymization methods should be applied to all training data to avoid privacy violations.

Acknowledgment. This work was supported by the JC STEM Lab of AI for Science and Engineering, funded by The Hong Kong Jockey Club Charities Trust, the Research Grants Council of Hong Kong (Project No. CUHK14213224).

References

- [Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.
- [Armando *et al.*, 2024] Matthieu Armando, Salma Galaaoui, Fabien Baradel, Thomas Lucas, Vincent Leroy, Romain Br  gier, Philippe Weinzaepfel, and Gr  gory Rogez. Cross-view and cross-pose completion for 3d human understanding. In *CVPR*, 2024.
- [Bian *et al.*, 2024] Siyuan Bian, Chenghao Xu, Yuliang Xiu, Artur Grigorev, Zhen Liu, Cewu Lu, Michael J Black, and Yao Feng. Chatgarment: Garment estimation, generation and editing via large language models. *arXiv:2412.17811*, 2024.
- [Blanz and Vetter, 2023] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries*. 2023.
- [Brooks *et al.*, 2023] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.
- [Cai *et al.*, 2024] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *NeurIPS*, 2024.
- [Chen *et al.*, 2022] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *TOG*, 2022.
- [Chen *et al.*, 2023a] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, et al. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *CVPR*, 2023.
- [Chen *et al.*, 2023b] Ziwei Chen, Qiang Li, Xiaofeng Wang, and Wankou Yang. Liftedcl: Lifting contrastive learning for human-centric perception. In *ICLR*, 2023.
- [Chen *et al.*, 2024] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. *arXiv:2412.10523*, 2024.
- [Ci *et al.*, 2023] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *CVPR*, 2023.
- [Dong *et al.*, 2023] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *CVPR*, 2023.
- [Dong *et al.*, 2024] Haiwei Dong, Yang Liu, Ted Chu, and Abdulmotaleb El Saddik. Bringing robots home: The rise of ai robots in consumer electronics. *arXiv:2403.14449*, 2024.
- [Feng and others, 2024] Yao Feng et al. Foundation Models for 3D Humans, 2024.
- [Feng *et al.*, 2024] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024.
- [Fu *et al.*, 2022] 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 *ECCV*, 2022.
- [Fu *et al.*, 2023] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, and Ziwei Liu. Unitedhuman: Harnessing multi-source data for high-resolution human generation. In *ICCV*, 2023.
- [He *et al.*, 2024] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *CVPR*, 2024.
- [Hong *et al.*, 2022] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. In *CVPR*, 2022.
- [Hu, 2024] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024.
- [Huang *et al.*, 2024a] Jie Huang, Ruibing Hou, Jiahe Zhao, Hong Chang, and Shiguang Shan. Refhcm: A unified model for referring perceptions in human-centric scenarios. *arXiv:2412.14643*, 2024.
- [Huang *et al.*, 2024b] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *ECCV*, 2024.
- [Jiang *et al.*, 2023] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2023.
- [Jin *et al.*, 2024] Sheng Jin, Shuhuai Li, Tong Li, Wentao Liu, Chen Qian, and Ping Luo. You only learn one query: learning unified human query for single-stage multi-person multi-task human-centric perception. In *ECCV*, 2024.
- [Ju *et al.*, 2023] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *CVPR*, 2023.
- [Juravsky *et al.*, 2024] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Superpadl: Scaling language-directed physics-based control with progressive supervised distillation. In *SIGGRAPH*, 2024.
- [Karras *et al.*, 2024] Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, et al. Fashion-vdm: Video diffusion model for virtual try-on. In *SIGGRAPH Asia*, 2024.
- [Khrodkar *et al.*, 2024] Rawal Khrodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024.
- [Kolotouros *et al.*, 2023] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai

- Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv:2306.09329*, 2023.
- [Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [Li *et al.*, 2024a] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *CVPR*, 2024.
- [Li *et al.*, 2024b] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *CVPR*, 2024.
- [Li *et al.*, 2024c] Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Unipose: A unified multi-modal framework for human pose comprehension, generation and editing. *arXiv:2411.16781*, 2024.
- [Lin *et al.*, 2023] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2023.
- [Lin *et al.*, 2024] Jing Lin, Yao Feng, Weiyang Liu, and Michael J Black. Chathuman: Language-driven 3d human understanding with retrieval-augmented tool reasoning. *arXiv:2405.04533*, 2024.
- [Lin *et al.*, 2025] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, and others. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models, 2025.
- [Liu *et al.*, 2023] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv:2310.08579*, 2023.
- [Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, et al. Visual instruction tuning. *NeurIPS*, 2024.
- [Luo *et al.*, 2024] Mingshuang Luo, Ruibing Hou, Hong Chang, Zimo Liu, et al. M3-gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *arXiv:2405.16273*, 2024.
- [Meng *et al.*, 2024] Yihang Meng, Hao Cheng, Zihua Wang, Hongyuan Zhu, Xiuxian Lao, and Yu Zhang. Efficient multi-modal human-centric contrastive pre-training with a pseudo body-structured prior. In *PRCV*, 2024.
- [Pan *et al.*, 2023] Xingang Pan, Ayush Tewari, Thomas Leimkühler, et al. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *SIGGRAPH*, 2023.
- [Shao *et al.*, 2024] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 360-degree human video generation with 4d diffusion transformer. *ACM Transactions on Graphics (TOG)*, 2024.
- [Tang *et al.*, 2023] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *CVPR*, 2023.
- [Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.
- [Wang *et al.*, 2023] Yizhou Wang, Yixuan Wu, Shixiang Tang, Weizhen He, Xun Guo, Feng Zhu, Lei Bai, Rui Zhao, Jian Wu, Tong He, et al. Hulk: A universal knowledge translator for human-centric tasks. *arXiv:2312.01697*, 2023.
- [Wang *et al.*, 2024a] Haoran Wang, Mohit Mendiratta, Christian Theobalt, and Adam Kortylewski. Facegpt: Self-supervised learning to chat about 3d human faces. *arXiv:2406.07163*, 2024.
- [Wang *et al.*, 2024b] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv:2410.21747*, 2024.
- [Wu *et al.*, 2023] Yue Wu, Sicheng Xu, Jianfeng Xiang, Fangyun Wei, Qifeng Chen, Jiaolong Yang, and Xin Tong. Aniportraitgan: animatable 3d portrait generation from 2d image collections. In *SIGGRAPH Asia*, 2023.
- [Wu *et al.*, 2024] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv:2405.17013*, 2024.
- [Xiong *et al.*, 2023] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, et al. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *ICCV*, 2023.
- [Xu *et al.*, 2024] Xinyu Xu, Yizheng Zhang, Yong-Lu Li, Lei Han, and Cewu Lu. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *arXiv:2406.19972*, 2024.
- [Yoshikawa *et al.*, 2023] Takato Yoshikawa, Yuki Endo, and Yoshihiro Kanamori. Stylehumanclip: Text-guided garment manipulation for stylegan-human. *arXiv:2305.16759*, 2023.
- [Yuan *et al.*, 2024] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, et al. Hap: Structure-aware masked image modeling for human-centric perception. *NeurIPS*, 2024.
- [Zhang *et al.*, 2024a] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *AAAI*, 2024.
- [Zhang *et al.*, 2024b] Yi Zhang, Wang Zeng, Sheng Jin, Chen Qian, Ping Luo, and Wentao Liu. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In *ECCV*, 2024.
- [Zhou *et al.*, 2024] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *CVPR*, 2024.