Chains of Diffusion Models

Yanheng Wei[1], Lianghua Huang[1*], Zhi-Fan Wu[1], Wei Wang[1], Yu Liu[1], Mingda Jia[2], and Shuailei Ma[3]

[1] Alibaba Group, Beijing, China yanheng.wyh@alibaba-inc.com
[2] Peking University Shenzhen Graduate School, Shenzhen, China
2201212832@stu.pku.edu.cn
[3] Northeastern University, Shenyang, China xiaomabufei@gmail.com

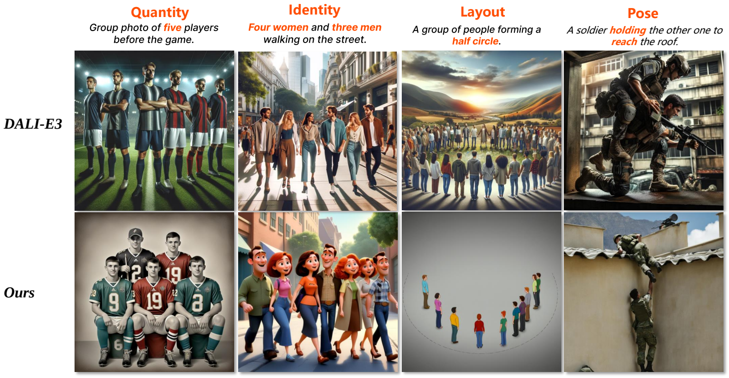

Quantity
Group photo of five players before the game.
Identity
Four women and three men walking on the street.
Layout
A group of people forming a half circle.
Pose
A soldier holding the other one to reach the roof.
DALI-E3
Ours

Fig. 1: Comparison of our *Chains* and *DALL-E3* [31], in multi-human scene generation. *Chains* outperform *DALL-E3* on the aspects of accurate human quantity, identity, layout, and pose with comparable image quality.

Abstract. Recent generative models excel in creating high-quality single-human images but fail in complex multi-human scenarios, failing to capture accurate structural details like quantities, identity accuracy, layouts and postures. We introduce a novel approach, **Chains**, which enhances initial text prompts into detailed human conditions using a step-by-step process. Chains utilize a series of condition nodes—text, quantity, layout, skeleton, and 3D mesh—each undergoing an independent diffusion process. This enables high-quality human generation and advanced scene layout management in diffusion models. We evaluate Chains against a new benchmark for complex multi-human scene synthesis, showing superior performance in human quality and scene accuracy over existing methods. Remarkably, Chains achieves this with under 0.45 seconds for a 20-step inference, demonstrating both effectiveness and efficiency.

Keywords: Diffusion model · Multi-human generation · Conditional generation

1 Introduction

The field of text-to-image synthesis has seen significant strides for the human generation in recent years, particularly in generating high-quality images for single-human [18, 26, 31, 32, 46] and simple human-centric scenes [2, 4, 35, 37, 38, 42, 44, 49]. However, synthesizing complex multi-human scenarios while ensuring image quality and structural accuracy (*e.g.,* quantities, layouts, and shapes) remains an open challenge [9, 22, 42, 44]. Generating complex scenes with multiple individuals demands several key aspects: accurately depicting the correct quantity of human instances amidst occlusion and intricate actions, preserving human identity according to the prompt, and ensuring the correct spatial distribution of multiple individuals. Moreover, it is also crucial that the appearance and pose of each human remain unambiguous.

However, existing general text-to-image methods, including advanced models such as GPT4-Vision, DALL-E3, and Gemini, encounter challenges in generating accurate representations of identity-bound multi-human scenes, particularly in complex scenarios where each person engages in different actions. These methods also exhibit shortcomings in detailing the overlap between multiple individuals, leading to face artifacts and deformations in human limbs and faces that are absent in single-person scenes. When it comes to the specific human-body generation or human-verb generation, they focus on single-human scenarios and often tend to generate ambiguous human crowds with wrong quantities and confounding body parts. As for the conditional generation methods like Layout-to-image and with other conditions, they only ensure the correctness of the spatial scene distribution, however lack high-quality human generation within overlapping conditions.

Contrary to these imperfect generative models, human designers typically adopt a sequential approach to such complex image creation. This process usually involves determining the layout, sketching the human structures, and then adding details. This method allows for the construction of scenes with complex spatial distributions in simpler, manageable steps, with each step governed by the outcome of its preceding steps.

Inspired by this process, we propose a *condition chains* paradigm for text-to-image multi-human synthesis, where we insert a sequence of condition *nodes* representing increasingly detailed human structures between the text and image. Each node in the chain conditions its predecessors for generation and acts as a *structural prior* to constrain the next node, culminating in the image generation.

Considering the structures of the human body, existing layout-conditioned text-to-image methods [12, 21, 48] inspire us to generate correct scene distributions correctly. Additionally, the use of keypoints [51] and 3D mesh [27, 33, 50] contributes to the creation of high-fidelity human figures. Hence we construct the condition chains with text, quantity, layout, skeleton, 3D mesh, and optional semantic embeddings for image generation. Each node is conditioned on all previous nodes, as depicted in Fig. 2. We utilize a transformer model [47] for each intermediate node to predict conditions, with a quantity prediction head and diffusion models [11] for synthesizing the following conditions from previous conditions. Specifically, each condition within Chains can be derived from preceding conditions, supporting the flexible combination of multiple conditional prompts beyond mere textual prompts.

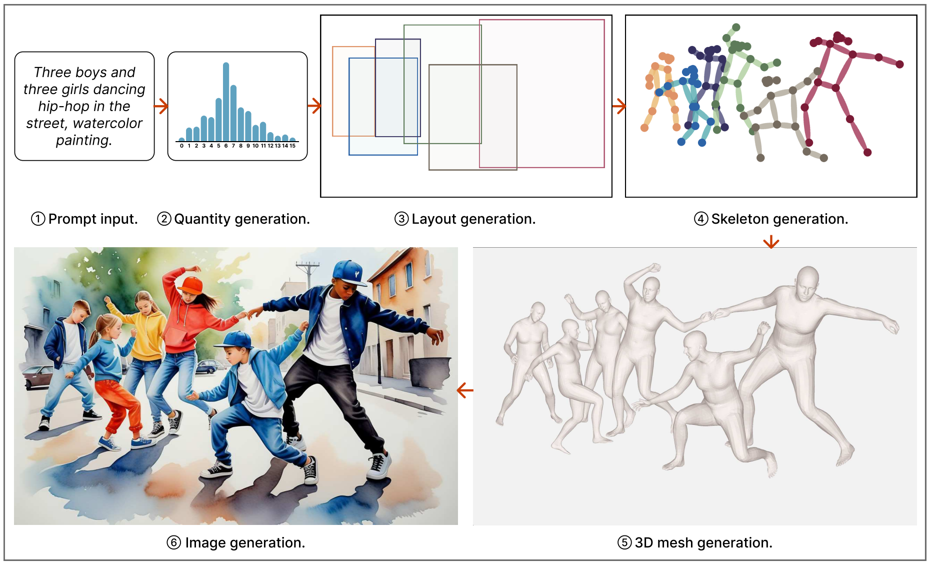


Fig. 2: Overview of the *chained generation* architecture. (a) Quantity prediction employs a transformer model for mapping text embeddings to softmax probabilities. (b) Layout, skeleton, 3D-mesh, and semantic embedding generations use transformer-based diffusion models for conditional denoising. (c) A modified ControlNet is utilized for generating the final image.

The final stage of image generation leverages a modified version of ControlNet, as detailed in [51], enriched with composable chain conditions. Benefiting from this flexible process, Chains are capable of generating synthetic intermediates at any point within the chain inference process. This characteristic equips Chains with both the capability of condition-to-multi-condition generation and multi-condition-to-image generation. For additional examples of generation and an overview of the architecture, we direct the reader to Fig. 1 and Fig. 2, respectively.

To assess the capability of generative models in producing complex multi-human scenes, we have created a benchmark comprising 86 prompts for multi-human scene generation, spanning several subcategories. These prompts encompass three splits including human quality, scene layouts, identity accuracy and human pose. Thus, providing a comprehensive evaluation of generative models under diverse structural constraints. Detailed information on the benchmark is available in Tab. 1.

Chains demonstrates its outperformance compared with recent state-of-the-art end-to-end generation methods in our spatial structure benchmark, highlighting the spatial modeling ability of Chains. What is more, Our method also shows better capability in attribute bindings compared to the end-to-end methods. Moreover, the generation speed of Chains is also fast, with each intermediate stage taking around 0.45 second to complete a 20-step generation. This enables users to swiftly control the intermediate generation results before proceeding to the generation of the final image. We have also conducted several ablation studies on the model and chain designs, indicating the potential for more efficient chain implementation through model scale reduction and multi-tasking.

2 Related Works

Text-to-image Diffusion Models. The field of text-to-image human generation has witnessed significant advancements. Diffusion models [2,4,11,34,35,37,38,42,44,45,49] have emerged as the dominant approach in both research and product communities, surpassing earlier works based on GANs [6,19,20,52] and autoregressive [30,39] models. However, the current text-to-image models [11,29,35,40] and human generation models [15,18,26] primarily excel in generating high-quality images for simple scenes with single or few humans. However, their overall performance degrades when dealing with complex scenes featuring multiple humans and objects. What is more, a notable challenge remains in accurately capturing and rendering the complex spatial relationships and ensuring the specific attributes of generated instances are consistent with the description in textual prompts.

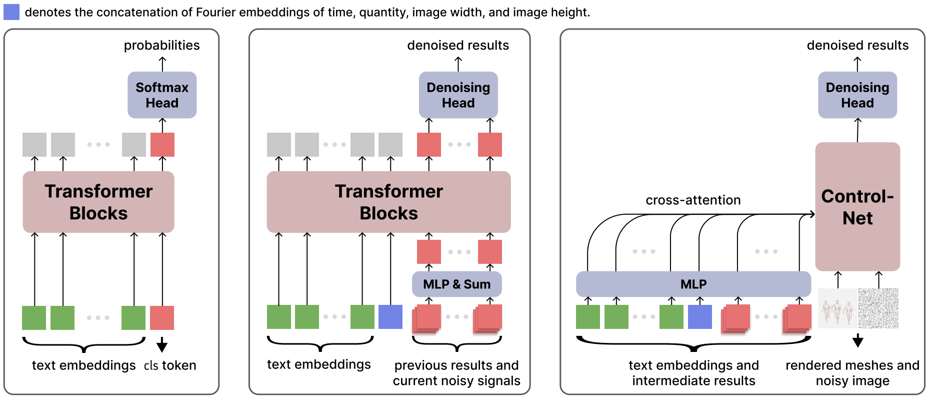
Layout-to-image Diffusion Models. Several recent works [7,12,21,48] aim to enhance generation accuracy under spatial structural constraints, focusing on aspects like object quantities, object location, layouts, and relationships. However, these methods often exhibit limited performance due to the due to the negative impact of explicit spatial constraints [21] and lack stability. Besides, they are fragile in generating high-quality human instances. In contrast, Chains applies smooth constraints to image generation and enhances the spatial condition at each step, which facilitates a balance between conditioning and high-quality output.

Multi-condition Diffusion Models. The methods most related to Chains are diffusion models with multi-instruction [9,14,41] along with ControlNet [51] that operate under multiple distinct conditions. However, the ControlNet have limited inter-dependencies between different conditions. The instruction paradigm requires instruction-specific tokens to aggregate different sub-task generators, which introduces redundancy. Different from them, our Chains could generate different conditioned intermediates within any stage of the forward progress of the condition chain, which is more flexible and benefits from the inter-conditions dependency.

3 Method

3.1 Overview

Our approach seeks to validate the proposed condition *chains* paradigm in the context of complex multi-human image generation. It decomposes the text-to-image generation process into a progressive conditional generation with condition chain sequence, starting from text and progressing through intermediate


🟦 denotes the concatenation of Fourier embeddings of time, quantity, image width, and image height.
probabilities
Softmax Head
Transformer Blocks
text embeddings
cls token
denoised results
Denoising Head
Transformer Blocks
MLP & Sum
text embeddings
previous results and current noisy signals
denoised results
Denoising Head
Control-Net
cross-attention
MLP
text embeddings and intermediate results
rendered meshes and noisy image

(a) Quantity prediction model. (b) Intermediate diffusion models. (c) ControlNet for image generation.

Fig. 3: Overview of the *chained generation* architecture. (a) Quantity prediction employs a transformer model for mapping text embeddings to softmax probabilities. (b) Layout, skeleton, 3D mesh, and semantic embedding generations use transformer-based diffusion models for conditional denoising. (c) A modified ControlNet is utilized for generating the final image.

nodes: quantity, layout, skeleton, 3D mesh, optional semantic embeddings, and ultimately to the final image. Each node serves as a *structural prior* for the generation of subsequent nodes. An overview of our approach is provided in Fig. 2

We utilize an independent transformer for each intermediate condition node in the chain. We model instance quantity prediction as a softmax classification task, while all other stages follow a denoising diffusion process. The final image generation employs a variant of ControlNet [51] with composable conditions. The following sections detail each step in the chain.

3.2 Intermediate Generation stages

Text Encoding. Text prompt serves as the initial input for the chain, conditioning the generation of subsequent nodes. We use the textual branch of the CLIP-ViT-G [8, 16, 36] model, with a moderate size of 662M, to obtain the embedding representation of the text. This model provides a text encoder aligned with visual content, which we expect to facilitate the prediction of the remaining nodes.

We exclude the last transformer block and the model's head, as they are more closely tied to the contrastive pretraining task of CLIP. Instead, we utilize the output of the penultimate layer as the text representation. We do not add layer normalization after the penultimate layer's output. Consequently, we encode a text into a tensor of dimensions $L \times 1280$, where L is the length of the tokenized text.

Category	Subcategory	Prompt
Quantity	Basic	*Five children walking on the street.*
Quantity	Calculation	*Two boys and three girls sitting on the grass.*
Quantity	Indefinite	*City hall hosts student sports competition.*
Layout	Basic	*Students sitting around the table.*
Layout	Basic	*A man standing close and the other standing far away.*
Layout	Relationship	*A soldier holding the other one to reach the roof.*
Layout	View angle	*Top view of people walking on the street.*
Layout	Arrangement	*A group of people forming a human pyramid.*
Layout	Arrangement	*Kids standing in a queue.*
Pose	Face	*A portrait of a little boy.*
Pose	Half-body	*A head-to-waist depiction of a girl.*
Pose	Full-body	*A young lady dances jazz in the street.*
Pose	Multi-individuals	*A football team training for the day before the match.*

Table 1: Categorized examples of prompts in our benchmark.

Quantity Generation. We train a softmax classifier on a large dataset of `(text,quantity)` pairs to model the text-conditional distribution of individual quantities in an image. We leverage a transformer to map text embeddings to softmax probabilities of quantity values, ranging from zero to a predefined maximum quantity. The transformer takes a concatenation of text embeddings and a learnable `cls` token as input, with the output derived from the `cls` token serving as logits. We utilize a cross-entropy loss function during training. During inference, we can adjust the sampling certainty of the quantity by applying a temperature parameter to the logits before calculating softmax probabilities.

Layout Generation. Layout generation involves predicting the location and size of each individual in an image. We express the layout as an $N \times 4$ tensor, where each row corresponds to an individual's bounding box. Each bounding box is described by a 4-dimensional vector `[cx,cy,w,h]`, where `[cx,cy]` are the center coordinates, and `[w,h]` represent the width and height of the bounding box. Each element of the vector `[cx,cy,w,h]` is normalized by dividing by the respective dimensions of the image width and height, ensuring that all values are within the range of `[0,1]`.

We model the layout generation as a denoising diffusion process conditioned on both the text embeddings and the predicted quantity. A transformer [47] is used to denoise the layout. The input of the transformer combines text embeddings, Fourier embeddings of the timestep, quantity, image width, image height, and embeddings of the noisy layout. We extract the embeddings of the noisy layout by applying a two-layer MLP with `SiLU` activation. We do not use positional embeddings as the individuals are unordered. The output from the last N tokens of the sequence, passed through a head layer, are taken as the denoised results.

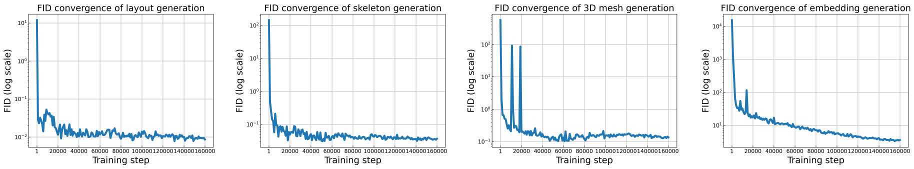


Fig. 5: Convergence curves for the FID metric across four stages (layout, skeleton, 3D mesh, and embedding generation) within our *chained generation* pipeline during the training process.

For batch-wise training or inference, we pad the bounding boxes to a maximum count, resulting in an $N_{\max} \times 4$ tensor, and use masked attention in the transformer to nullify the impact of padded tokens.

Skeleton Generation. The skeleton generation stage predicts the pose of each individual, represented as a 17-keypoint representation from the COCO-Pose dataset [25]. We use YOLOv8 [17] to gather pose data, treating detections as the groundtruth. Consequently, we express the skeletons of individuals in an image as an $N \times (17 \times 3)$ tensor. Each keypoint is represented by a 3-dimensional `[x,y,score]` vector, with `[x,y]` denoting the coordinates and `score` the prediction confidence. We normalize `[x,y]` by dividing them by image width and height, respectively, ensuring a value range of `[0,1]`.

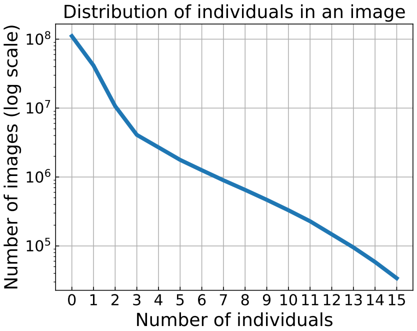


Fig. 4: The distribution of the number of individuals per image in the training data, which reveals a significant imbalance.

Like the layout distribution, we use a denoising diffusion model to learn the skeleton distribution in an image. A transformer maps noisy skeletons to denoised ones, leveraging the same architecture as in the layout generation, with the input embeddings replacing the noisy layout embeddings by the point-wise addition of clean layout and noisy skeleton embeddings. We derive the skeleton embeddings through a two-layer MLP with `SiLU` activation. The denoising results are obtained from the output of the last N tokens of the sequence passed through a head layer.

3D mesh Generation. The 3D mesh generation stage predicts a SMPLX [33], which is a comprehensive 3D model that encapsulates the body, fully articulated hands, and expressive face, for each human, providing a refined structure of the body. This representation is compactly encoded as an 182-dimensional vector.

Hence, an image's 3D models can be collectively represented as an $N \times 182$ tensor.

We use a denoising diffusion model to learn the distribution of 3D models of individuals in an image. The transformer architecture is the same as the layout and skeleton generation modules, with the input embeddings replacing the last N tokens with the point-wise addition of clean layout, clean skeleton, and noisy 3D model embeddings. The 3D model embeddings are derived through a two-layer MLP with `SiLU` activation. The denoising results are obtained from the output of the last N tokens passed through a head layer.

3.3 Semantic Embedding Generation

In this stage, an embedding vector is synthesized to represent each individual's attributes. To obtain the *groundtruth* embedding vector, we extract the features from the facial region of each individual using the visual branch of the CLIP-ViT-G [8,16,36] model. We directly use the global output from the `cls` token of the CLIP model as the representation. As a result, the individual embeddings in an image can be expressed as an $N \times 1280$ tensor. We apply L2 normalization on the tensor and multiply the output by $\sqrt{1280}$ to ensure a standard deviation of around 1.0.

Similar to the generation of layout, skeletons, and 3D meshs, we use a denoising diffusion model with a transformer architecture to learn the embedding distribution. The transformer is similar to that described in layout, skeleton and 3D mesh generation in Sec. 3.2, except that we replace the last N tokens of the input embeddings with the point-wise addition of clean layout embeddings, clean skeleton embeddings, clean SMPL embeddings, and noisy semantic embeddings. We use a two-layer MLP with `SiLU` activation to map the semantic embeddings before adding them with other conditions. We obtain the denoising results by projecting the output of the last N tokens through a head layer.

3.4 Image Generation

The last stage in the generative chain is the image generation. This is modeled using a variant of ControlNet [51] with composable conditions. Specifically, we consider three conditions: the SMPLs rendered on a black background, the SMPL-X parameters, and the semantic embeddings. The meshes are rendered to match the size of the target image and are then processed by ControlNet through a hint embedding module equipped with stacked convolutional blocks. This serves to map the conditions to additive feature maps. In scenarios involving multiple individuals, each 3D mesh is visualized and fed separately into the hint embeddings. The results are then summed before being passed to the remaining modules of ControlNet. The SMPL-X parameters and semantic embeddings are mapped through separate MLP layers to align the dimension with text embeddings. The mapping results are then pointwisely summed and concatenated with text embeddings along the sequence length dimension, which are subsequently used as the cross-attention context for ControlNet. We randomly

Methods	Quantity	Identity	Layout	Pose	Average
SDXL [35]	0.60	0.63	0.61	0.59	0.60
SDXL variant [35]	0.63	0.69	0.70	0.65	0.66
ControlNet w/ SDXL variant [51]	0.78	0.65	0.67	0.66	0.69
Midjourney	0.80	0.69	0.71	0.73	0.73
Ours	**0.83**	**0.75**	**0.78**	**0.73**	**0.77**

Table 2: Average human ratings for text-to-image generation models, computed by averaging scores within categories and then across categories for overall score.

zero-out the mapped semantic embeddings with a probability of 50% to support image generation both with and without semantic embeddings as the condition.

4 Experiments

4.1 Data Preparation

We gather a comprehensive text-image dataset from publicly available sources including LAION-5B [43], COYO-700M [3], CC12M [5], and DataComp-1B [10]. To ensure high quality and appropriate content, we meticulously applied quality and content filters, resulting in a refined collection of 200 million text-image pairs.

In the next step of data preparation, we utilize the YOLOv8 model [17] to analyze each image. This model help us determine the number of individuals present, their respective bounding boxes, and pose skeletons. Additionally, we employ the OSX model [23] to estimate the SMPL-X parameters [33] for each individual depicted in the images. The frequency distribution of individuals per image is illustrated in Fig. 4.

Upon analyzing the dataset, we identify a significant skew in the distribution of individuals per image. To counteract this imbalance and ensure robust model performance across different numbers of individuals, we adopt a balanced sampling approach. We categorize the dataset based on the number of individuals in each image. During the training process, we sample from each category i with a probability $p_i \sim M_i^\tau$ proportional to the category's size M_i raised to the power of τ. We empirically set $\tau = 0.5$ across all training stages.

Metric	SDXL variant [35]	Ours
FID	5.00	5.31

Table 3: FID scores of our *chained generation* framework compared to the end-to-end baseline on the subset of COCO validation dataset where at least one person is present in the image.

For the training of the quantity prediction model, we utilized the entire dataset. In contrast, for the training of generative models of layout, skeletons, 3D meshes, semantic embeddings, and images, we use a subset of the dataset featuring at least one individual per image.

4.2 Implementation Details

Each intermediate generation stage (*i.e.,* the generation of *quantity, layout, skeletons, 3D meshes* and *semantic embeddings*) utilizes an independent transformer model [47] with approximately 1 billion parameters. The transformer has 24 layers with a dimension of 2048 and 32 attention heads. We employ the AdamW optimizer [28] with a learning rate of 1×10^{-4}, a weight decay of 0.1, and a maximum gradient norm of 4.0 for training the quantity prediction model. The layout, skeleton, 3D mesh, and semantic embedding generation models are trained with a learning rate of 1×10^{-4} and a weight decay of 0.06. We use a diffusion process with a cosine schedule and v-prediction mode [13, 24]. Each stage undergoes a training period of 160,000 steps, conducted on 16 A100 GPUs, utilizing a total batch size of 4096.

For the image generation phase, we utilize a variant of SDXL [1, 35] that excels at portrait generation as the base model. On this base model, we train the ControlNet [51], incorporating composable conditions including SMPL-X parameters, optional semantic embeddings, and rendered meshes. This training adheres to the original diffusion parameters of the SDXL model. Here, we set the learning rate at 1×10^{-6} and do not apply weight decay.

Evaluation aspects	Quantity	Layout	Pose	Overall
Success rate	0.87	0.82	0.78	0.82

Table 4: Average human ratings for text-to-3D mesh generation, computed by averaging scores within categories and then across categories for the overall score.

4.3 Benchmark

We develop a human prompt benchmark to evaluate the ability of generative models to synthesize intricate scenes while complying with structural constraints. The benchmark, comprising 86 manually crafted prompts across categories such as quantity, viewpoint, spatial layout, relationship, action, and posture, enables a comprehensive assessment of generative models under diverse conditions. The prompts are concise and clear, facilitating rapid assessment of the generated results. Using these intricate constraints to evaluate generative models sheds light on their capabilities and flaws. Tab. 1 illustrates the benchmark's examples.

Generation task	Layout	Skeleton	3D mesh
Separate models	0.0086	0.037	0.135
Multi-task model	0.0132	0.042	0.152

Table 5: Ablation analysis on multi-task *v.s.* separate models for layout, skeleton, and 3D mesh generation.

4.4 Chains-to-Image Generation

In this section, we evaluate the quality of the generated images and their alignment with corresponding query prompts. To measure the image quality, we employ the Fréchet Inception Distance (FID) metric, comparing the generated images to real images from a subset of the COCO dataset [25]. This subset includes images with at least one person present in their annotations. The FID scores are presented in Tab. 3. Our approach achieves a comparable FID score with the end-to-end baseline.

To assess the alignment between the generated images and the query prompts, we involve human evaluators, with three annotators assigned to each query, and their responses are averaged. Leveraging the prompt benchmark detailed in Sec. 4.3, we conduct a comparative analysis between the end-to-end baseline and our chained generation approach regarding adherence to structural constraints. The voting results, both in terms of individual aspects and overall average, are outlined in Tab. 2. For the performance of baselines on our Chains benchmark, we feed the text prompts directly to SDXL, SDXL variant and Midjourney, as for the ContrloNet, we utilize both the text prompt and Chains-generated human keypoints as the conditions for a fair comparison.

The experiment results highlight the robust spatial structure understanding and representation ability of Chians, in all four evaluation splits. Chains outperform the base method ControlNet with the SDXL variant by an average 8% enhancement, and especially 10% and 11% gain on the Identity and Layout. Compared with the recent sota Midjourney, Chains also has better spatial ability and an average 4% higher performance. Notably, our chained generation approach surpasses the end-to-end baseline in terms of adhering to structural constraints.

4.5 Text-to-Mesh Generation

To better understand the complete text-to-image chain, we initially explore the transition from text to 3D mesh generation. These SMPLs offer crucial insights into structural compliance, independent of image generation. This separation prevents premature engagement with complex details.

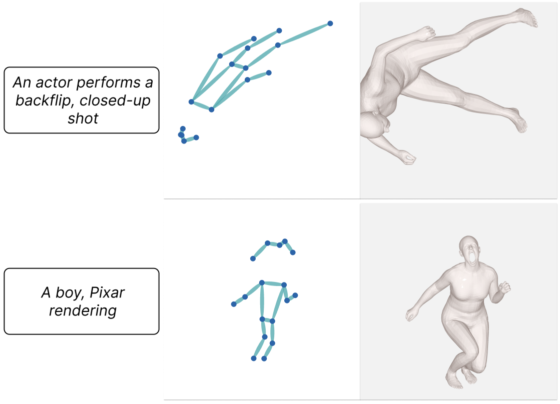


Fig. 6: Failure cases of text-to-mesh generation. The mesh generation model fails to depict uncommon poses and counterfactual human body proportions, such as cartoon-like figures.

FID and Convergence: We establish a validation dataset containing 10,000 unique text-image pairs, ensuring no overlap with the training data. For performance evaluation, we compute the Fréchet Inception Distance (FID), comparing structures generated from these texts with the ground truth derived from the validation images. Throughout the training,

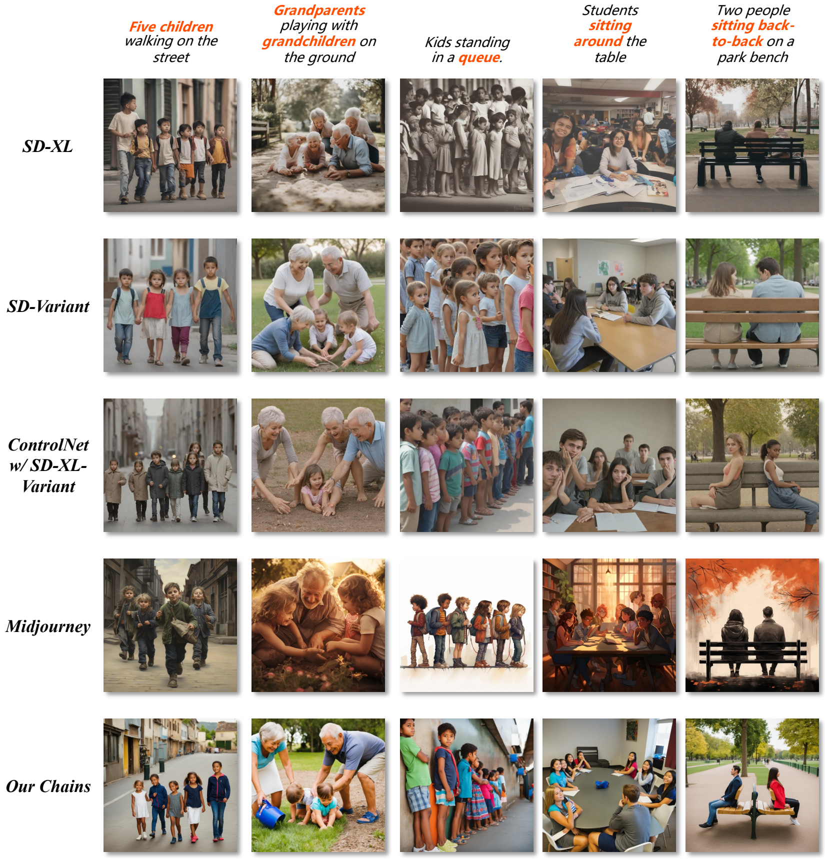

Five children walking on the street
Grandparents playing with grandchildren on the ground
Kids standing in a queue.
Students sitting around the table
Two people sitting back-to-back on a park bench
SD-XL
SD-Variant
ControlNet w/ SD-XL-Variant
Midjourney
Our Chains

Fig. 7: More generation comparisons between the state-of-the-art methods. For instance, in the first column, only our **Chains** generates the accurate number, and in the fourth column, the rest of the models do not accurately focus on the layout and pose, with the detailed analysis presented in section.4.6.

we monitor FID, plotting convergence curves for layout, skeleton, 3D mesh, and embedding generation stages in Fig. 4, demonstrating the models' promising statistical performance and training stability.

Compliance with constraints: To assess how well the generated structures adhere to the specified constraints, we rely on human evaluators, as this aspect is challenging to assess via automated algorithms or models. For the benchmark detailed in Sec. 4.3, we assign three annotators to each query. They judge whether the generated structures conform to the prompts, with their responses averaged to obtain a final score. Results for each constraint category, as well as aggregated voting outcomes, are presented in Tab. 4. Overall, the generated structures demonstrated a high degree of compliance with the prompts, especially for constraints related to quantity.

4.6 Visual Comparison with State-of-the-Arts

We provide an in-depth visual comparison in Fig. 7, where each column showcases images generated by baseline methods and our Chains, responding to prompts that focus on various elements of image structure. As demonstrated in the first column, only Chains accurately produces an image featuring $five$ children, showcasing its superior capability in determining the correct number of human figures. In the second column, the initial three methods do not successfully create images of grandparent pairs, and Midjourney produces only a single grandchild. Conversely, Chains precisely generates an image with the correct human identities as per the prompt, ensuring the number of figures is accurate. Notably, in the fourth column, only Chains effectively synthesizes a group of students sitting *around* a round table, as opposed to incorrectly placing them on one or two sides of a rectangular table, highlighting Chains' adeptness at understanding spatial relationships. In the final column, Chains alone accurately generates two people positioned back-to-back, whereas all other baseline methods fail to replicate this scenario.

4.7 Ablation Analysis

In this section, we conduct a comprehensive analysis of our chained generation framework, exploring various aspects and configurations.

Multi-tasking: We explore the performance of merging the layout, skeleton, and SMPL generation tasks into a single model. In this setup, the three tasks are distributed across different GPU nodes and jointly trained with a shared transformer backbone. Results are detailed in Tab. 5 and compared to the single-task baseline. While we observe a speedup, there is also a reduction in performance. We acknowledge the trade-off and leave the exploration of multi-tasking for efficiency improvements in future endeavors.

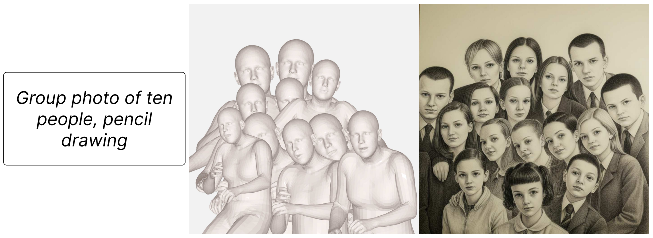


Fig. 8: A failure case of image generation. The image generation may exhibit misalignment with the SMPL, especially when the number of individuals is large.

Failure cases of SMPL generation: Fig. 6 illustrates some of the failure cases related to SMPL generation. These issues are primarily attributed to misalignment, possibly stemming from inaccuracies in the estimation of the 3D mesh model when generating groundtruth meshes. SMPL-X may encounter challenges in representing meshes for uncommon poses (*e.g.,* backflips), even when skeletons are accurately generated. Counterfactual human body proportions, such as cartoon-like figures with oversized heads and slender limbs, also pose challenges.

Failure cases of image generation: Fig. 8 showcases an instance where image generation encounters difficulties. In scenarios involving numerous individuals, the generated images may exhibit misalignment with the SMPLs, sometimes introducing additional persons even when the conditioned SMPLs do not contain them. These challenges highlight room for potential future improvements.

5 Conclusion and Discussion

This paper introduces a novel generation framework for high-quality multi-human scene generation with multi-modal condition chains, breaking down the complex task of generation with high-fidelity human conditions into a condition evolution sequence of several simple manageable stages. Our framework, demonstrated through comprehensive benchmarking, outperforms traditional end-to-end text-to-image methods and layout-to-image methods with better human quality and more precise spatial layout distributions. The ability to manipulate intermediate stages further enhances the human interoperable and instruction usage of Chains.

Our approach also faces limitations, such as misalignments and inaccuracies in mesh and image generation, particularly in challenging poses or crowd scenes with even large numbers of humans and objects. What is more, the implementation of the chain condition approach for objects remains unexplored. Future work will focus on addressing these limitations, and also adapting our framework to a wider range of domains.

References

1. Stoked Reality XL. `https://civitai.com/models/146039/stoked-reality-xl?modelVersionId=162531`, [Accessed 24-09-2023]
2. Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., et al.: Cm3: A causal masked multimodal model of the internet. arXiv preprint arXiv:2201.07520 (2022)
3. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
4. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
6. Chen, Y., Li, G., Jin, C., Liu, S., Li, T.: Ssd-gan: Measuring the realness in the spatial and spectral domains. In: AAAI (2021)
7. Cheng, J., Liang, X., Shi, X., He, T., Xiao, T., Li, M.: Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation (2023)

8. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
9. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
10. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Hoe, J.T., Jiang, X., Chan, C.S., Tan, Y.P., Hu, W.: Interactdiffusion: Interaction control in text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
13. Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., Dittadi, A.: Diffusion models for video prediction and infilling. arXiv preprint arXiv:2206.07696 (2022)
14. Hu, H., Chan, K.C.K., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., Chang, M.W., Jia, X.: Instruct-imagen: Image generation with multi-modal instruction (2024)
15. Huang, S., Gong, B., Feng, Y., Chen, X., Fu, Y., Liu, Y., Wang, D.: Learning disentangled identifiers for action-customized text-to-image generation (2023)
16. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773
17. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), https://github.com/ultralytics/ultralytics
18. Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: HumanSD: A native skeleton-guided diffusion model for human image generation (2023)
19. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis (2023), https://arxiv.org/abs/2303.05511
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
21. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. CVPR (2023)
22. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)
23. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21159–21168 (2023)
24. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. arXiv preprint arXiv:2305.08891 (2023)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
26. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. arXiv preprint arXiv:2310.08579 (2023)

27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
30. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018)
31. OpenAI: Dall·e 3: Image generation model. `https://openai.com/dall-e-3` (2023), accessed: yyyy-mm-dd
32. OpenAI: Gpt-4: Enhancements and capabilities. `https://openai.com/blog/gpt-4` (2023), accessed: yyyy-mm-dd
33. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
34. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
35. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
38. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
39. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering (2021)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
41. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2023)
42. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
43. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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)
44. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of captions. arXiv preprint arXiv:2006.11807 (2020)

45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022)
46. Team, G.: Gemini: A family of highly capable multimodal models (2023)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
48. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7452–7461 (2023)
49. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
50. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV) (2020)
51. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
52. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks (2020)

Chains of Diffusion Models
Supplementary Material

Yanheng Wei[1], Lianghua Huang[1]*, Zhi-Fan Wu[1], Wei Wang[1], Yu Liu[1], Mingda Jia[2], and Shuailei Ma[3]

[1] Alibaba Group, Beijing, China yanheng.wyh@alibaba-inc.com
[2] Peking University Shenzhen Graduate School, Shenzhen, China
2201212832@stu.pku.edu.cn
[3]

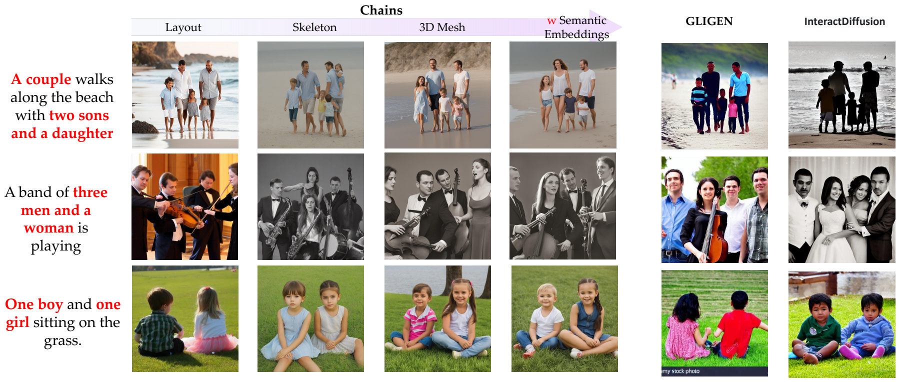

Chains
Layout
Skeleton
3D Mesh
w Semantic Embeddings
GLIGEN
InteractDiffusion
A couple walks along the beach with two sons and a daughter
A band of three men and a woman is playing
One boy and one girl sitting on the grass.

Fig. 1: Visualization for **Chains**, we meticulously ablate each component within **Chains** through generated cases, including **layout**, **skeleton**, **3D mesh**, and **semantic embeddings**. Additionally, we compare our results with the current state-of-the-art layout to image method, GLIGEN [21] and InteractDiffusion [12], where the sota models use the same layout condition as our **Chains**.

A Closer Visual Comparisons

In addition to the comparative analysis presented in our submission, we have included further comparisons with state-of-the-art methods such as GLIGEN [21] and InteractDiffusion [12] in Fig. 1.

B Ablation for Chains Components.

In Fig. 1, we conduct the visual comparison of the component-to-image generation using **Chains**. It is evident that the layout condition provides customized supervision of multi-human distribution and locations, which is shared by recent layout-to-image methods. Additionally, the inclusion of skeleton and 3D mesh components introduces more fine-grained supervision, enhancing the generation of high-quality human details such as clear faces, hands, and fingers. This comparison indicates that the chain-of-thought generation method outperforms traditional text-to-image and layout-to-image methods in producing high-quality multi-human scenes.

C Model Limitations

C.1 Conditions out of control.

As illustrated in Tab. 2 and Tab. 4 of our paper, our model demonstrates robust structural control in the majority of cases. The rare occasional instances of suboptimal control, as observed in Fig. 6, can be attributed to the inaccuracies in crowd detection and counting. The discrepancies in Fig. 7 stem from the out-of-distribution of the SMPL-X normal shapes. We believe that adopting more accurate representations could mitigate these issues and further improve the control accuracy. Thanks.

C.2 Discussion on error accumulation.

In the chained generation process, errors are typically stage-specific and orthogonal, meaning that issues like quantity and layout are confined to individual stages and do not propagate through subsequent stages. Moreover, **Chains** is inherently robust, for example, it is capable of generating accurate meshes from imperfect skeletons, effectively correcting initial deviations. As demonstrated in Fig. 2, the experiments confirm that the sequential error accumulation has a limited impact on our model's performance.

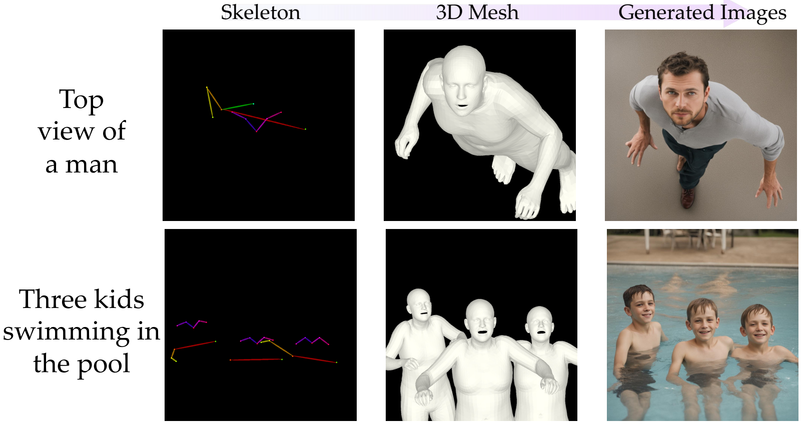


Fig. 2: This figure shows that 3D mesh prediction is error-tolerant to skeleton condition prediction.

C.3 Limited control over hand posture and facial expression.

We found that, the 3D mesh sometimes provides limited control over hand and face, it is mainly due to the inaccuracies of OSX algorithm in estimating 3D meshes for these delicate areas. These inaccuracies can lead to misalignments in the mesh-to-image generation model. Using a more advanced model like SMPLer-X could improve control and quality.