SKDream: Controllable Multi-view and 3D Generation with Arbitrary Skeletons

Yuanyou Xu[1], Zongxin Yang[2], Yi Yang[1*]
[1]ReLER, CCAI, Zhejiang University
[2]DBMI, HMS, Harvard University
{yoxu,yangyics}@zju.edu.cn, zongxin_yang@hms.harvard.edu

Abstract

Controllable generation has achieved substantial progress in both 2D and 3D domains, yet current conditional generation methods still face limitations in describing detailed shape structures. Skeletons can effectively represent and describe object anatomy and pose. Unfortunately, past studies are often limited to human skeletons. In this work, we generalize skeletal conditioned generation to arbitrary structures. First, we design a reliable mesh skeletonization pipeline to generate a large-scale mesh-skeleton paired dataset. Based on the dataset, a multi-view and 3D generation pipeline is built. We propose to represent 3D skeletons by Coordinate Color Encoding as 2D conditional images. A Skeletal Correlation Module is designed to extract global skeletal features for condition injection. After multi-view images are generated, 3D assets can be obtained by incorporating a large reconstruction model, followed by a UV texture refinement stage. As a result, our method achieves instant generation of multi-view and 3D contents that are aligned with given skeletons. The proposed techniques largely improve the object-skeleton alignment and generation quality. Project page at https://skdream3d.github.io/.

1. Introduction

In view of visual representation dimension [64, 72], 2D image generation [16, 53, 57], multi-view (2.5D) generation [35, 55], and 3D generation [17, 27, 69] have advanced successively and achieved significant progress. Given the success of large language models [23, 26, 43, 48–50], textual representation has been widely applied in generation [51–53], as in other domains [13, 14, 73]. To achieve more controllable generation, conditions beyond text have attracted considerable attention. 2D image conditions (*e.g.*, edge maps, human skeletons, and concept references) [54, 75] have been well studied. Similarly in 3D generation, 2D [30] and 3D [10] conditions have also been explored.

Although the aforementioned conditions in controllable generation complement text descriptions, they still struggle to precisely describe shape structures. In contrast, skeletons, among various types of conditions, exhibit superior ability to depict shape structures: ***(i) Representation of object anatomy.*** Skeletons can efficiently represent various 3D structures with sparse joints and bones, while other conditions like depth maps fail to represent full 3D anatomy. ***(ii) Articulation into different poses.*** Skeletons are widely used for character animation in computer graphics [3, 22] due to their simplicity and efficiency. Other conditions such as rough shapes [10] are inconvenient for pose articulation. ***(iii) Freedom of editing.*** Given an initial skeleton, users can freely add new structures or modify joints and bones to create ideal shapes. Examples are in Fig. 1.

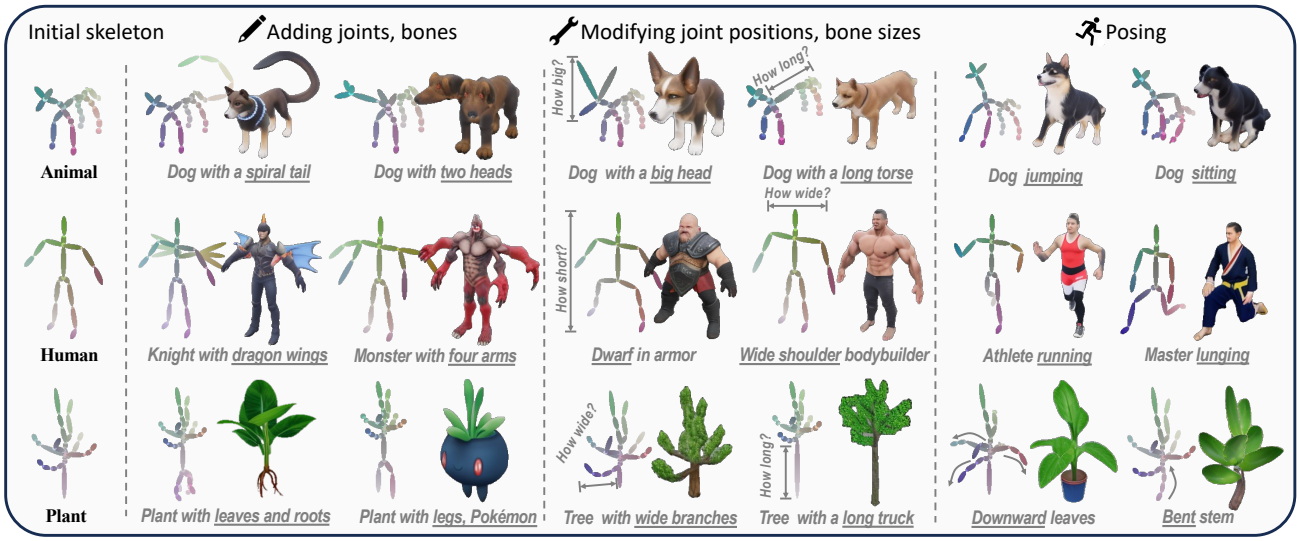
Despite these advantages, previous studies [19, 20, 40, 75, 77] on conditional generation are limited to human skeletons. From the perspective of generalization, we would like to ask: ***Is it possible to use arbitrary skeletons as conditions to generate any creatures or even general objects?***

To achieve this goal, we identify two main challenges that hinder the use of arbitrary skeletal conditions for generation: ***(i) Lack of large-scale object-skeleton pairs for training***. Extensive studies [4, 15, 37] on 2D/3D human pose estimation have made human-skeleton paired data readily available. However, when skeletal structures are unknown, estimating arbitrary skeletons from 2D images or videos becomes challenging due to its ill-posedness. ***(ii) Insufficiency of 2D information to describe arbitrary skeletons***. Human skeletons are relatively simple and can be described by a fixed set of 2D joints. However, complex skeletons suffer from self-occlusion and ambiguity, necessitating 3D information to fully capture their anatomy and pose.

To address these challenges, we focus on multi-view and 3D generation with skeletal conditions. To tackle data scarcity problem, we construct ***a large-scale dataset Objaverse-SK containing mesh-skeleton pairs.*** Textured meshes are selected from Objaverse [8] by semantic classes to form a subset. In order to achieve reliable mesh skele-

*Corresponding Author

(a) Text and skeletons for **Collaborative Appearance and Shape Controlling**



(b) Skeleton-based editing for **Accurate Anatomy and Pose Controlling**

Figure 1. **Demonstration of skeletal conditions for controllable generation.** We argue that ***skeletons and text provide complementary description for shape and appearance respectively***, as shown in (a). Moreover, ***flexible and accurate control of object anatomy and pose can be realized by editing the joints and bones in skeletons***, as shown in (b). Arbitrary skeletal structures are supported in our framework. Multiple views are generated and only front-view images are shown.

tonization, we propose a new pipeline to generate skeletons with sparse joints from meshes. The pipeline mainly consists of curve skeleton extraction and curve simplification, achieving an 80% success rate, significantly outperforming previous deep learning based method RigNet [71].

To fully control object anatomy and pose, we build the ***skeletal conditioned generation model in a multi-view manner.*** We represent a 3D skeleton with conditional skeleton images by *Coordinate Color Encoding (CCE)* to reduce ambiguity. Joints and bones are encoded with unique colors according to their 3D positions. For condition injection, we designed a *Skeletal Correlation Module (SCM)* to extract features from these conditional images and then generate multi-view images for the object. Later, a Large Reconstruction Model (LRM) is employed to produce 3D assets from the multi-view images. To address potential blurriness during reconstruction, we enhance appearance quality by a refinement stage that up-samples the multi-view images to higher resolutions and refines the texture in UV space.

The experimental results indicate that our framework achieves instant generation of multi-view and 3D contents which are aligned with given skeletons. The proposed coordinate color encoding and the skeletal correlation module significantly improve the object-skeleton alignment, and accelerate model convergence by $5\times$. 3D assets conforming to the given skeleton can be generated in $\sim$20s and refined in $\sim$60s. To the best of our knowledge, this work is a pioneer in achieving arbitrary skeletal conditioned generation with following contributions:

- Constructing **the first large-scale dataset**, Objaverse-SK, containing mesh and skeleton pairs that cover diverse skeletal structures. We developed a pipeline to generate sparse skeletons from meshes with a high success rate.
- Proposing a **multi-view and 3D generation pipeline** for arbitrary skeletons, including *coordinate color encoding* for compact condition representation and the *skeletal correlation module* for effective condition injection.

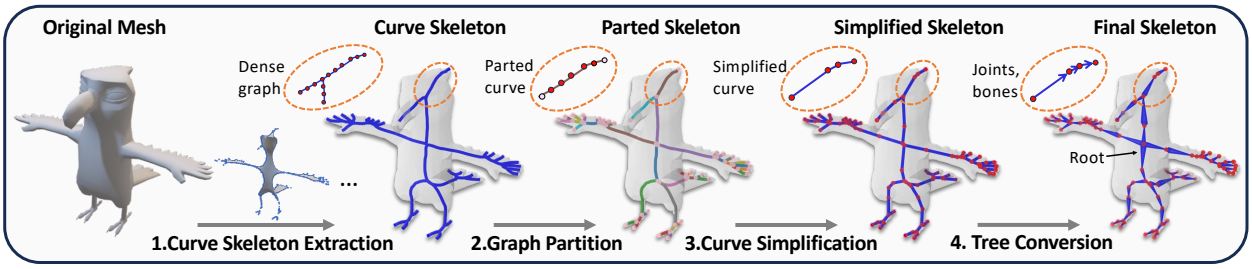


Figure 2. **Illustration of the pipeline for skeleton generation from meshes** (§3.2). The curve skeleton is first extracted from the given mesh, followed by simplification of parted curves. The curve graph is converted to a tree as the final skeleton.

2. Related Work

Controllable 2D Generation. Based on image diffusion models [51–53], versatile controlling conditions have been studied. In terms of spatial controlling, ControlNet [75] and other similar works [40, 78] train a side network for spatial conditions such as edge maps, normal maps and human skeletons. Some works focus on human image generation from skeletons [18, 20, 62]. Box-based instance controlling is also concerned in some works [29, 79, 81]. As for content controlling, DreamBooth [54] finetunes the model to bind the given subject with an identifier in text prompt. IP-Adapter [74] trains an adapter to inject styles or concepts into the model. Some works [31, 32, 63] also focus on human ID control. Besides, some methods [2, 38, 39, 41] can achieve conditional generation without fine-tuning.

Controllable 3D Generation. Controlling contents in 3D generation can be easily achieved by controlling image-to-3D generation, which has been studied by plenty of works [17, 27, 59, 69]. However, in the image-to-3D paradigm, spatial controlling for 3D generation is not as easy as content controlling. Many works enhance controlling ability upon score distillation paradigm [28, 46, 83]. Coin3D [10] presents a framework to control the multi-view diffusion and 3D generation by shape proxies, i.e. combination of simple basic shapes. Sculpt3D [6] enhances text-to-3D generation with retrieved 3D priors. Sherpa3D [34] proposes to generate a coarse shape with a 3D diffusion model and refine the shape with SDS [46]. Some works for 3D human or head avatar generation [19, 24, 33, 77, 80, 82] uses human skeleton or facial landmarks as the condition in 2D or 3D space. Recent 3D native diffusion models [7, 21, 25, 67, 76] indicate promising results. Clay [76] designs a transformer-based [45, 60] 3D diffusion framework and various conditions like images and point clouds can be injected through cross-attention layers. A recent work MVControl [30] realizes 3D generation with single-view 2D spatial conditions like normal maps and edge maps by conditional multi-view generation and 3D reconstruction. Our work shares the similar workflow, but we focus on general skeleton conditioned generation, which has never been studied by previous works. Another recent work AnimatableDreamer [65] generates 4D objects with skeletons extracted from given videos by canonical score distillation, while our work develops a totally different framework for instant generation. As for skeletal representation, AnimatableDreamer extracts dense joints and unstructured bones from videos, while our method creates sparse joints and structured bones, which are more similar to human-made skeletons and more efficient for shape representation and animation.

Mesh Skeletonization. Various algorithms were designed for extracting skeletons from 3D meshes. [58] and [1] compute curve skeletons (C-S) via iterative mesh contraction operations. [11, 66] proposed to extract skeletons medial axis transformation skeleton (MAT-S) by point selection and connection prediction. C-S and MAT-S can serve as shape representation, while human-made skeletons (H-S) are often different from them. Since the main purpose is animation, H-S only contain sparse joints and bones. Some works [70, 71] propose data-driven approaches to learn mesh skeletonization from human annotated data. In this work, we have tried learning-based method [71] but found the results were not satisfactory. Therefore, we develop a new pipeline to generate skeletons which are as sparse as H-S while keep the shape of C-S.

3. Dataset Construction

3.1. Data Preparation

The largest existing open dataset containing mesh-skeleton pairs is ModelResources [70]. There are around 3,000 3D meshes without textures. The scale is insufficient to train a text-driven generative model, and it lacks textures for appearance modeling. To address these limitations, we construct a dataset with 8× larger scale and includes color textures. Our dataset, named Objaverse-SK, is built upon a large-scale 3D dataset Objaverse [8]. Although our data generation pipeline is applicable to a broad range of object categories, we focus on three main categories including "Animals", "Human Shapes" and "Plants", as they can typically be represented by tree-structured skeletons. Category labels are obtained from G-Objaverse [47]. Consequently, our dataset contains 24k 3D meshes, consisting of 15k animals, 6k human shapes and 3k plants. Text prompts for

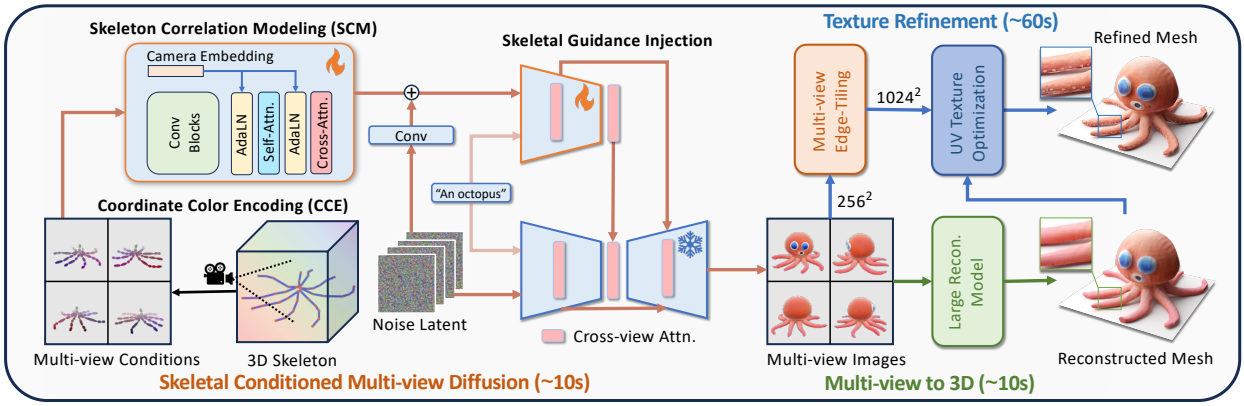


Figure 3. **Illustration of the pipeline for skeletal conditioned multi-view and 3D generation** (§4). The 3D skeleton is projected into 2D images and represented by coordinate color encoding. The skeletal correlation is modeled by skeletal correlation module, and then is injected into the diffusion model. Multi-view images are first generated and then a 3D textured mesh is reconstructed. The texture is further refined through UV-space optimization. Our framework achieves instant and high-quality generation given arbitrary skeletons.

these models are generated by Cap3D [36].

3.2. Skeleton Generation

To obtain mesh-skeleton pairs, a effective method for generating skeletons from meshes is crucial. There are two concerns: the skeleton structure and success rate. The skeleton structure should accurately describe the object anatomy and be suitable for posing. Furthermore, an ideal method should generate reasonable skeleton structures with a high success rate. We tested a learning-based method RigNet [71] (Fig. 4). Although the generated skeleton structures can be close to human-made skeletons in its training data, it primarily produces symmetric skeletons and tends to be unstable across diverse anatomies and poses.

Skeleton extraction. To enhance flexibility and robustness, we design a new reliable pipeline, utilizing curve skeletons as the intermediate representation. An illustration of the pipeline is in Fig. 2. Given the structural inconsistency between curve skeletons and human-made skeletons, we further convert dense curves into sparse joints and bones. The detailed pipeline is elaborated below. 1) Initially, Mean Curvature Flow (MCF) [58] is employed to generate curve skeletons from meshes robustly. 2) Next, we build a graph from the set of curves generated from the mesh, consisting of dense nodes and edges. Intersection nodes (degree$>$1) are identified and the graph is divided into several parts by these nodes. 3) In each part, the curve contains no branches so it can be simplified by the Douglas-Peucker algorithm (DP) [12] into line segments.

Tree conversion. At this stage, the basic shape of the skeleton is established, but the root position and the bone directions between joints still need to be determined. The problem can be regarded as a graph-to-tree conversion. First, a spanning tree is constructed from the graph to eliminate cycles. We then identify high-degree intersection nodes as candidates for the root. To ensure an efficient structure, the skeleton is configured by selecting the tree with the minimum height among these candidates. This approach ensures that the root node is located at a significant intersection, minimizing the distances between the root and other joints. More details can be found in the appendix.

4. Generation Pipeline

4.1. Skeletal Conditioned Multi-view Generation

As the dataset is constructed, we consider building the conditional generative model based on it. Since unconditional multi-view diffusion models have been well studied, we start directly from a base model MVDream [55] and focus on the conditional generation. Two main issues are of concern: i) how the skeleton is represented, and ii) how it is injected into the model.

Skeletal condition representation. As we aim to generate images which align with the given skeletons, using spatial guidance in the diffusion model is a reasonable approach. Skeletons can be projected from 3D space to the image plane as 2D conditions. However, overlapping and information loss occur during the projection, which may lead to semantic and structural ambiguities for spatial guidance, as illustrated in Fig. 8. Thus, incorporating richer information is crucial to mitigate these ambiguities.

Coordinate Color Encoding (CCE). In order to preserve 3D information, we encode joint coordinates using spatial colors. While prior works [28, 61] use canonical color maps for shape representation, our approach focuses on representing skeletons with sparse joints and bones. We begin by normalizing skeletons within a canonical cube $[0, 1]^3$. Each position in this cube corresponds to a unique color, with RGB values precisely matching the positional coordinates. As a result, the 2D conditional image can represent

the 3D spatial positions of the skeleton. For bones, we assign colors based on their midpoint. Additionally, we incorporate normalized values of view-dependent inverse depth of the skeleton as the alpha channel (CCE-D). With the absolute spatial coordinates and relative depth encoded in the conditional images, there will be more precise and richer guidance information for generation.

Skeletal condition injection. Spatial conditions such as canny edges and normal maps have been investigated in 2D image diffusion models. In ControlNet [75], the conditional image is encoded by convolution blocks, resulting in an output spatial size that matches the latent size. Then, the condition features are added to the latent features. The encoder of the original diffusion model is copied as a side network to produce guidance features, which are fused with the original features in the decoder. Our pipeline adopts this paradigm from ControlNet, and further enhances it with a more effective condition feature extraction module.

Skeletal Correlation Modeling (SCM). For a skeleton in 3D space, we first project it into multi-view images as 2D conditions. Given the sparse nature of skeletons in spatial dimension, convolution blocks lack global modeling capacity. To address this, we design a Skeletal Correlation Module (SCM) to enhance the condition features by modeling the anatomical correlation among different parts of a skeleton, and the view correlation for different projection views. The structure of the module is in Fig. 3. ***(i) First, anatomical correlation is extracted by a self-attention layer***, which constructs the global skeleton features for each view. ***(ii) Then, the cross-view correlation is modeled by a cross-attention layer***, allowing the extraction of correspondences among skeleton images from multiple views. This enables the model to recognize identical joints in different views. In addition, we use adaptive layer normalization [68] to fuse the camera pose embedding with the skeletal features. Associating each skeleton image with a camera pose helps to generate view-dependent shapes. Adding correlation layers during condition encoding significantly facilitates learning, achieving 5× faster convergence (Fig. 9).

4.2. Multi-view Images to 3D Generation

Instant reconstruction. Given the generated multi-view images, we use a Large Reconstruction Model (LRM), specifically InstantMesh [69] for fast textured mesh reconstruction. However, the reconstructed textures often appear blurry. On one hand, the resolution of generated images is 256^2, which struggles to capture fine details. On the other hand, the appearance quality also degrades during reconstruction. In order to recover and further enrich the appearance, we introduce a new refinement stage.

Appearance refinement. First, the generated multi-view images are upscaled 4 times into 1024^2 by Stable Diffusion with ControlNet-Tile [75]. ControlNet-Edge [9, 75] is used to maintain the shape consistency across different views during tiling. Once tiled, these high-resolution images I_i^h are used to refine the reconstructed texture. A learnable 2D texture u in UV space is created and initialized as the reconstructed texture u_0, and then images are rendered through differentiable rendering $\mathcal{R}(u, c_i)$ for given camera views c_i. The MSE loss is optimized between the rendered images and tiled high-res images. Moreover, a regularization term is added to maintain consistency in UV space:

$$\mathcal{L}_u = \sum_i ||I_i^h - \mathcal{R}(u, c_i)||_2^2 + \lambda * ||u - u_0||_2^2. \quad (1)$$

Since the high-res images cannot cover every position on the mesh, some regions of u will not be optimized, e.g. bottom of the object. We found these regions are unstable during optimization and may produce unexpected artifacts (see Fig. 10). The regularization term helps the optimized texture maintain the appearance from u_0 in these regions. Consequently, the high-frequency details can be learned in covered regions while the global consistency can also be achieved in uncovered regions. The optimization can be finished within 15 seconds.

4.3. Object-Skeleton Alignment Evaluation

Contrastive alignment. In order to measure how much an object is aligned with a skeleton, we develop a new evaluator, named Contrastive Object-Skeleton Alignment (COSA). We use the self-supervised DINOv2 [44] as the backbone F to extract both object and skeleton features. Then, the alignment adapter G_θ consisting of several self-attention layers is used to modulate the features. The adapter ends with an average pooling layer to aggregate the aligned features into a vector. Similar to CLIP [49], we train the adapter using contrastive learning by InfoNCE loss [42, 56]. Finally, the skeleton alignment score (SKA) can be calculated by cosine similarity between the features from an object image x and a skeleton image y as:

$$\mathcal{S}_{\text{SKA}}(x, y) = \cos(G(F(x)), G(F(y))). \quad (2)$$

COSA guided diffusion. Based on COSA, another conditional generation pipeline can also be realized, following the approach proposed in [2]. On each denoising time step t, the approximate clean image $\hat{x}_0$ is estimated from the predicted noise ϵ_t as in DDIM [57]. The estimated clean image and skeleton condition are fed into COSA to calculate the alignment loss $\mathcal{L}_{\text{COSA}}(\hat{x}_0, y) = 1 - \mathcal{S}_{\text{SKA}}(\hat{x}_0, y)$. Then the predicted noise is modified by the gradient of the alignment loss for actual denoising:

$$\hat{\epsilon}_t = \epsilon_t + s(t) \cdot \nabla \mathcal{L}_{\text{COSA}}(\hat{x}_0, y) \quad (3)$$

where $s(t)$ controls the guidance strength. With the additional guidance of the alignment loss, the generated object will tend to follow the conditional skeleton y.

Method\SKA Score	Training	Mean$_{Inst.}$	Mean$_{Class}$	Animals	Humans	Plants	Apodes	Bipeds	Quadrupeds	Arthropods	Wings
SDEdit [38]	❍	70.13	65.06	79.50	64.70	50.99	74.74	75.24	82.73	79.07	79.21
SDEdit+COSAG	◗	72.11	67.32	80.91	67.60	53.46	79.73	83.17	85.98	77.80	73.82
Ours-Raw	●	74.69	67.09	88.39	69.90	42.97	**94.74**	82.14	90.18	89.81	84.69
Ours-SCM	●	**80.43**	**74.38**	**91.16**	**78.45**	**53.53**	94.47	**85.20**	**94.19**	**90.68**	**88.40**

Table 1. **Quantitative comparison of object-skeleton alignment (SKA) score** (§5.2). Alignment scores are calculated over three classes (blue) and five subclasses of animal (green). The average scores over all instances and three classes (pink) are also shown. The highest scores among all methods are bold and the highest scores among baseline methods are underlined.

Method	Training	PickScore				CLIP Score			
		Win Rate	Animals	Human	Plants	Mean$_{Inst.}$	Animals	Human	Plants
SDEdit [38]	❍	19.56	18.23	14.66	28.98	29.04	29.47	28.71	**28.29**
SDEdit+COSAG	◗	18.17	18.54	19.95	15.06	28.99	29.52	28.50	28.11
Ours-Raw	●	28.99	29.06	28.37	**29.55**	**29.90**	30.10	**30.02**	**29.23**
Ours-SCM	●	**33.28**	**34.17**	**37.02**	26.42	29.83	**30.23**	29.84	28.75

Table 2. **Quantitative comparison of PickScore and CLIP Score** (§5.2). Scores are calculated over three classes (blue) and averaged over all instances (red). The highest scores among all methods are bold and the highest scores among baseline methods are underlined.

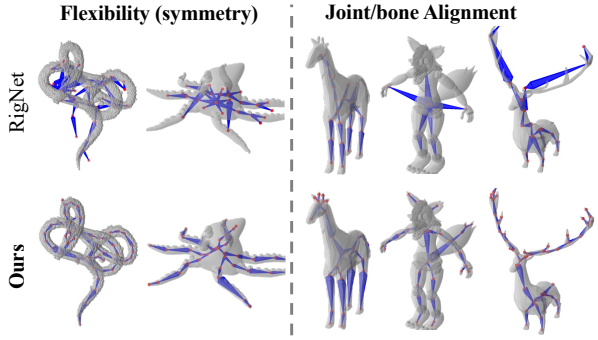


Figure 4. **Comparison of skeletons generated** from 3D meshes by RigNet [71] and our method (§5.1).

5. Experiment

5.1. Results of Mesh Skeletonization

We compare our method with the learning-based method RigNet [71], and the results are shown in Fig. 4. RigNet tends to produce symmetric skeletons with limited flexibility and misaligned joints/bones, resulting in an around 15% success rate. In contrast, our method supports arbitrary-pose skeletons and achieves better joint/bone alignment. It produces more reliable results with an 80% success rate. More details and results can be found in appendix.

5.2. Results of Multi-view Generation

Evaluation protocols. The Objaverse-SK evaluation set contains 108 skeleton instances, covering three main classes: animals, human shapes and plants. As animals include diverse skeleton structures, we further divide them into more detailed subclasses (examples are shown behind): Apodes (fish, snakes), Bipeds (ducks, penguins), Quadrupeds (dogs, bears), Arthropods (scorpions, crabs), Wings (birds, dragons). Three evaluation metrics are considered for multi-view generation: SKA Score for skeletal alignment, PickScore for image quality, and CLIP Score for textual alignment. Furthermore, samples from categories excluded in training are obtained from ShapeNet [5] for generalization evaluation (Fig. 7). Additionally, evaluation results on human-made skeletons are in appendix.

Baseline methods. Since there is no prior work that can achieve arbitrary skeletal conditioned generation, we implement two methods for comparison. The first baseline is SDEdit [38]. The process starts from adding noise on conditional images with a time step (set as 0.7). Then clean images are generated by denoising steps. The method is entirely unsupervised. The second baseline is the COSA Guidance (COSAG) derived from [2], which is elaborated in Section 4.3. The guidance strength is set as $s(t) = 7.5\sqrt{1-\alpha_t}$. Since we found it cannot achieve stable results, it is combined with SDEdit. The method requires an extra model so it is half supervised. Ours is fully supervised on object-skeleton pairs.

Qualitative comparison. The qualitative results are shown in Fig. 5. Given skeleton images as conditions, SDEdit can produce images following the skeleton. However, limited by the editing capacity, the generated objects often have incorrect anatomies. For example, the snake body is broken, and the donkey body is generated as wooden. When it is enhanced by the COSAG, the quality of generated content is improved in some cases but still unsatisfactory. Compared with them, our results show superior quality and are more consistent with both skeletal and textual conditions.

Quantitative comparison. Comparison results of skeleton alignment are shown in Table 1. SDEdit-based methods have around 70 SKA scores, while ours can achieve 80. Among three classes, animals tend to have higher alignment scores while plants have lower scores. Since plants

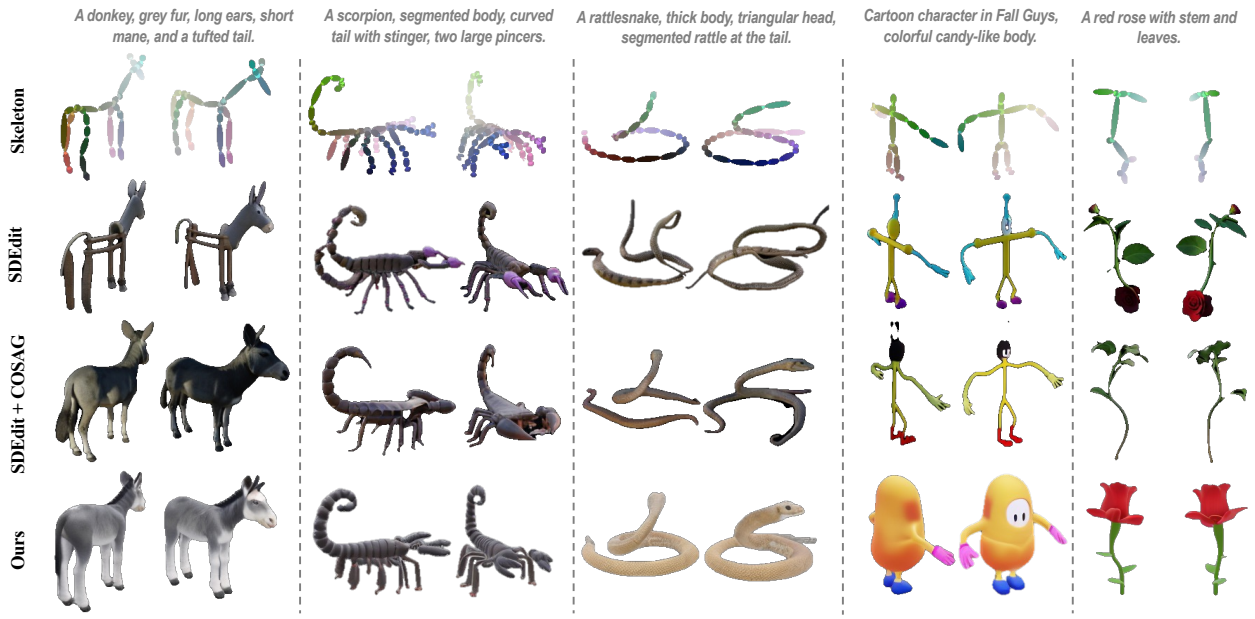


Figure 5. **Qualitative comparison of skeletal conditioned multi-view generation** (§5.2). Conditional skeletons and text prompts are shown above. Four views are generated and two views are shown for simplicity.

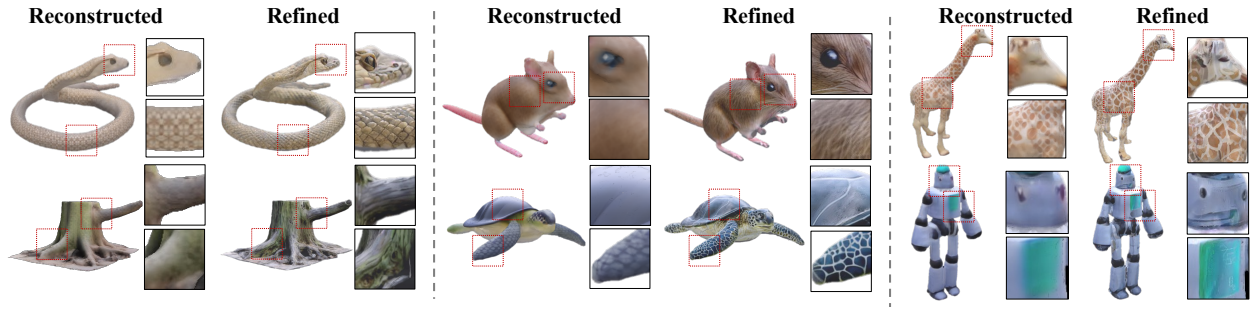


Figure 6. **Qualitative comparison of textured meshes before and after refinement** (§5.3). Rendered color images are shown. Local areas are enlarged for better viewing.

may have more complex structures and sometimes extend further from skeletons. For five sub-classes, our method achieves consistently high alignment scores. With the help of SCM and CCE-D skeletons, both the alignment scores and pick scores are further improved.

5.3. Results of 3D Generation

Texture refinement. Results of 3D reconstruction from multi-view images are shown in Fig. 6. The raw reconstructed and refined results are compared. The raw textures are blurry and lack details, while the proposed refinement stage can significantly enhance texture quality.

Rigging and animation. Given a motion sequence of a skeleton, our method can be applied to generate 4D animation. Textured mesh can be generated given the skeleton at the rest pose, and then directly be rigged and skinned for animation. Demo videos can be found in project page[1].

[1] https://skdream3d.github.io/.

6. Ablation Study

Skeletal condition representation. The skeletal condition representation we use consists of coordinate color encoding (CCE) with depth alpha (D). The ablation results are shown in Fig. 8 and Tab. 3. Richer information in conditions can help the model determine the content better. As a result, better alignment can be achieved. In Fig. 8 right, the skeleton of a penguin is highly ambiguous. If CCE-D is used, the body pose and orientation of the penguin can be successfully inferred from colors.

Skeletal correlation modeling. We show the effect of our skeletal correlation module in Tab. 3 and Fig. 9. SCM achieves better alignment scores than convolutional blocks. SCM with layer normalization (LN) achieves a 4× faster convergence rate. Furthermore, if LN is replaced with the adaptive LN (AdaLN), the model can achieve an SKA score of 70 within 1k training steps. The results indicate that extracting global features from conditional images is crucial.

Figure 7. **Qualitative results of novel categories in ShapeNet [5]** (§5.2), which are not covered by the training set of Objaverse-SK.

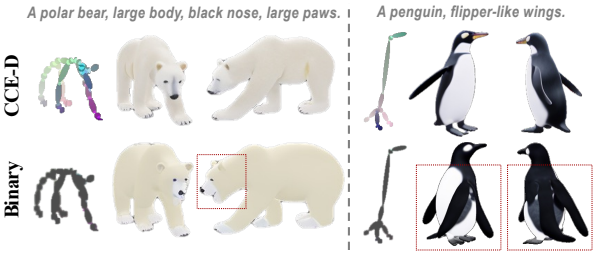


Figure 8. **Ablation study of coordinate color encoding with depth alpha (CCE-D)** (§6). Richer information can help the model to avoid ambiguity and generate better anatomy.

Module	Skeleton	SKA Score	Animals	Humans	Plants
SCM	CCE-D	**80.43**	**91.16**	**78.45**	**53.53**
SCM	CCE	78.97	90.45	74.82	52.57
SCM	Binary	77.27	89.20	76.04	46.16
Conv	CCE-D	76.65	88.53	75.04	46.17
Conv	Binary	74.69	88.39	69.90	42.97

Table 3. **Ablation study of skeletal module and representation types** (§6). SCM and CCE-D achieve higher alignment scores.

Refinement regularization. We show the ablation results of appearance refinement in Fig. 10. The refined appearance contains rich details such as snake scales and wood grain, in comparison to the reconstructed results. However, artifacts also appear in the regions that are not covered by high-res images. With the help of UV space regularization, the artifacts are effectively removed in uncovered regions. As a result, consistent colors are maintained from original textures and details are enhanced during optimization.

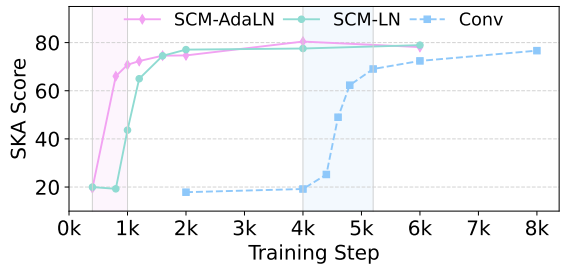


Figure 9. **Convergence processes of different skeletal modules** (§6). SCM with AdaLN achieves 5x faster convergence.

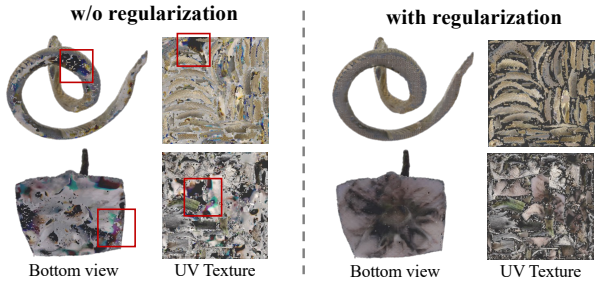


Figure 10. **Ablation study of UV space regularization** (§6). Bottom views and UV textures are shown. Front views of the snake and the tree stump can be found in the first column of Fig. 6.

7. Limitation and Future Work

Although our work achieves arbitrary skeletal conditioned generation, there are still many problems that can be further studied. The skeletons we currently use may have limited descriptive ability for non-tree structured objects. More powerful shape representations can be studied as new conditions. In addition, our work only considers global skeletons without fine-grained semantics. Injecting detailed semantics into the skeleton parts is a meaningful topic to study. For future work, the skeleton condition can be studied in native 3D generation frameworks. Moreover, our dataset may also be used in other tasks such as arbitrary skeleton estimation from images. More discussion is in the appendix.

8. Conclusion

In this work, we propose to use skeletons as the structural condition for controllable generation. First, we construct a large-scale 3D mesh-skeleton paired dataset. We propose an effective mesh skeletonization method to generate mesh-aligned sparse skeletons with a high success rate. Based on the dataset, we present a skeletal conditioned multi-view generation pipeline. Coordinate color encoding and skeletal correlation module are proposed to realize efficient condition representation and injection. Furthermore, 3D meshes can be instantly reconstructed, followed by a refinement stage to achieve better texture quality. In summary, our work achieves controllable multi-view and 3D generation with arbitrary skeletons as conditions.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62293554), and in part by the Natural Science Foundation of Zhejiang Province (LDT23F02023F02) and the Fundamental Research Funds for the Zhejiang Provincial Universities (226-2024-00208).

References

[1] Andreas Bærentzen and Eva Rotenberg. Skeletonization via local separators. *ACM Transactions on Graphics (TOG)*, 40 (5):1–18, 2021. 3

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 3, 5, 6

[3] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007. 1

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6, 8

[6] Cheng Chen, Xiaofeng Yang, Fan Yang, Chengzeng Feng, Zhoujie Fu, Chuan-Sheng Foo, Guosheng Lin, and Fayao Liu. Sculpt3d: Multi-view consistent text-to-3d generation with sparse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10228–10237, 2024. 3

[7] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. *arXiv preprint arXiv:2409.12957*, 2024. 3

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1, 3

[9] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 5

[10] Wenqi Dong, Bangbang Yang, Lin Ma, Xiao Liu, Liyuan Cui, Hujun Bao, Yuewen Ma, and Zhaopeng Cui. Coin3d: Controllable and interactive 3d assets generation with proxy-guided conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 1, 3

[11] Zhiyang Dou, Cheng Lin, Rui Xu, Lei Yang, Shiqing Xin, Taku Komura, and Wenping Wang. Coverage axis: Inner point selection for 3d shape skeletonization. In *Computer Graphics Forum*, pages 419–432. Wiley Online Library, 2022. 3

[12] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10 (2):112–122, 1973. 4

[13] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation instruction generation with bev perception and large language models. In *ECCV*, 2024. 1

[14] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Scene map-based prompt tuning for navigation instruction generation. In *CVPR*, 2025. 1

[15] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022. 1

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[17] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1, 3

[18] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[19] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 1, 3

[21] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[22] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46, 2007. 1

[23] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1

[24] Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, et al. Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. *arXiv preprint arXiv:2312.03763*, 2023. 3

[25] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy.

Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pages 112–130. Springer, 2024. 3

[26] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*, 2023. 1

[27] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 1, 3

[28] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 3, 4

[29] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[30] Zhiqi Li, Yiming Chen, Lingzhe Zhao, and Peidong Liu. Mvcontrol: Adding conditional control to multi-view diffusion for controllable text-to-3d generation. *arXiv preprint arXiv:2311.14494*, 2023. 1, 3

[31] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 3

[32] Chao Liang, Fan Ma, Linchao Zhu, Yingying Deng, and Yi Yang. Caphuman: Capture your moments in parallel universes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6400–6409, 2024. 3

[33] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, pages 1508–1519. IEEE, 2024. 3

[34] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20763–20774, 2024. 3

[35] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1

[36] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[37] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 1

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 6

[39] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 3

[40] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1, 3

[41] Marianna Ohanyan, Hayk Manukyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Zero-painter: Training-free layout control for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8764–8774, 2024. 3

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[43] OpenAI. Gpt-4 technical report, 2023. 1

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

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

[47] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024. 3

[48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020. 1

[51] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 3

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[54] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 3

[55] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 4

[56] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 5

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 5

[58] Andrea Tagliasacchi, Ibraheem Alhashim, Matt Olson, and Hao Zhang. Mean curvature skeletons. In *Computer Graphics Forum*, pages 1735–1744. Wiley Online Library, 2012. 3, 4

[59] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3

[60] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[61] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 4

[62] Jiajun Wang, Morteza Ghahremani, Yitong Li, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. *arXiv preprint arXiv:2406.02485*, 2024. 3

[63] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[64] Wenguan Wang, Yi Yang, and Yunhe Pan. Visual knowledge in the big model era: Retrospect and prospect. *Frontiers of Information Technology & Electronic Engineering*, 26(1):1–19, 2025. 1

[65] Xinzhou Wang, Yikai Wang, Junliang Ye, Fuchun Sun, Zhengyi Wang, Ling Wang, Pengkun Liu, Kai Sun, Xintong Wang, Wende Xie, et al. Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. In *European Conference on Computer Vision*, pages 321–339. Springer, 2024. 3

[66] Zimeng Wang, Zhiyang Dou, Rui Xu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Shiqing Xin, Taku Komura, Xiaoming Yuan, and Wenping Wang. Coverage axis++: Efficient inner point selection for 3d shape skeletonization. In *Computer Graphics Forum*, page e15143. Wiley Online Library, 2024. 3

[67] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3

[68] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019. 5

[69] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 3, 5

[70] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, and Karan Singh. Predicting animation skeletons for 3d articulated models via volumetric nets. In *2019 international conference on 3D vision (3DV)*, pages 298–307. IEEE, 2019. 3

[71] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*, 2020. 2, 3, 4, 6

[72] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 1

[73] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). *arXiv preprint arXiv:2401.08392*, 2024. 1

[74] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 3, 5

[76] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3

[77] Muxin Zhang, Qiao Feng, Zhuo Su, Chao Wen, Zhou Xue, and Kun Li. Joint2human: High-quality 3d human generation via compact spherical embedding of 3d joints. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1429–1438, 2024. 1, 3

[78] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[79] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 3

[80] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21057–21067, 2023. 3

[81] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. 3

[82] Zhenglin Zhou, Fan Ma, Hehe Fan, Zongxin Yang, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2024. 3

[83] Wenjie Zhuo, Fan Ma, Hehe Fan, and Yi Yang. Vividdreamer: invariant score distillation for hyper-realistic text-to-3d generation. In *European Conference on Computer Vision*, pages 122–139. Springer, 2024. 3