SK-Adapter: Skeleton-Based Structural Control for Native 3D Generation

Anbang Wang[1⋆], Yuzhuo Ao[1⋆], Shangzhe Wu[2†], and Chi-Keung Tang[1†]

[1] The Hong Kong University of Science and Technology
[2] University of Cambridge
{awangas,yaoaa}@connect.ust.hk, sw2181@cam.ac.uk, cktang@cs.ust.hk

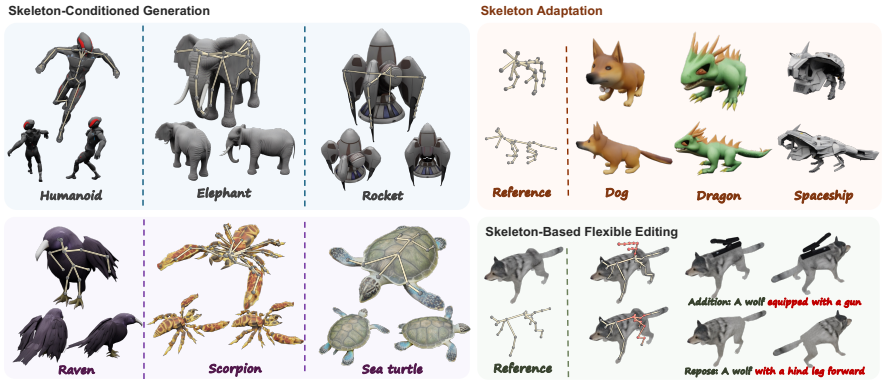

Fig. 1: SK-Adapter efficiently generates 3D assets in native 3D domain from given skeletons with its lightweight and effective designs, which also supports skeleton adaptation and flexible skeleton-based editing.

Abstract. Native 3D generative models have achieved remarkable fidelity and speed, yet they suffer from a critical limitation: inability to prescribe precise structural articulations, where precise structural control within the native 3D space remains underexplored. This paper proposes SK-Adapter, a simple yet efficient and effective framework that unlocks precise skeletal manipulation for native 3D generation. Moving beyond text or image prompts, which can be ambiguous for precise structure, we treat the 3D skeleton as a first-class control signal. SK-Adapter is a lightweight structural adapter network that encodes joint coordinates and topology into learnable tokens, which are injected into the frozen 3D generation backbone via cross-attention. This design allows the model to not only effectively "attend" to specific 3D structural constraints but also preserve its original generative priors. To bridge the data gap, we contribute the Objaverse-TMS dataset, a large-scale

⋆ Equal contribution. † Equal advising.

dataset of 24k text-mesh-skeleton pairs. Extensive experiments confirm that our method achieves robust structural control while preserving the geometry and texture quality of the foundation model, significantly outperforming existing baselines. Furthermore, we extend this capability to local 3D editing, enabling region-specific editing of existing assets with skeletal guidance, which is unattainable by previous methods. Project page: `https://sk-adapter.github.io/`

Keywords: Skeleton-based Control · 3D Generation · Adapter

1 Introduction

The field of 3D content creation is undergoing a paradigm shift, transitioning from time-consuming optimization-based methods [35, 48] to efficient, feed-forward native 3D generative models [10, 21, 53–55, 65]. These emerging frameworks, typically built upon scalable flow transformers, can synthesize high-fidelity 3D assets from text or images in mere seconds. However, despite their impressive visual quality and speed, precise structural controllability remains a formidable challenge. While text prompts convey high-level semantics and image prompts provide view-specific visual cues, both fall short in prescribing precise 3D articulations for entire assets, such as "bending the knee 60 degrees", or defining atypical anatomical topologies. For generated assets to be usable in animation and gaming pipelines, explicit structural control is indispensable.

Among various structural representations like bounding boxes or point clouds, skeletons are standard, compact, and hierarchical abstractions for articulated objects and are widely used for downstream tasks like animation. In the realm of 2D generation, the integration of skeletal guidance has achieved remarkable success. Pioneering works such as ControlNet [67] and T2I-Adapter [31] demonstrated that injecting 2D human pose images into diffusion models allows for precise control over the layout and posture of generated images. Inspired by this success, recent 3D approaches like SKDream [56] have attempted to replicate this paradigm to 3D generation by adopting a "2D lifting" strategy. They project arbitrary 3D skeletons into 2D projections to condition multi-view diffusion models, which are subsequently lifted to 3D via reconstruction.

However, directly transferring this 2D-conditioned paradigm to 3D generation introduces a fundamental dimensionality mismatch. While 2D skeletons work well for 2D images, compressing intrinsic 3D structures into 2D planes for 3D generation inevitably leads to spatial ambiguity—depth information is flattened, and self-occlusions in projected views cause the model to misinterpret complex topologies. Furthermore, relying on multi-stage reconstruction pipelines often degrades texture quality and introduces geometric artifacts, limiting the fidelity of the final assets.

To achieve precise structural fidelity without compromising generation quality, we posit that the control signal must be congruent with the generation space. Skeletal information should be injected directly within the native 3D domain, bypassing the lossy 2D projection bottleneck. However, realizing this goal faces

the difficulty of adapting large-scale 3D transformers to follow strict skeletal structure constraints without catastrophic forgetting of their generative priors.

In this work, we propose **SK-Adapter**, a novel framework for efficient, skeleton-guided native 3D generation. Our key insight is that precise structural control stems from the strict spatial alignment between the guidance signal and the generative latents. Instead of forcing the model to interpret abstract features or foreign 2D projections, we conceptualize the 3D skeleton as a set of sparse spatial tokens. These tokens encapsulate both geometric coordinates and topological constraints, which are then seamlessly injected into the transformer backbone through novel skeletal cross-attention layers, allowing the model to attend to precise 3D spatial constraints during volumetric generation.

Adapter methods, which originated in natural language processing and are now being quickly adopted for customizing pretrained transformer models by inserting lightweight modules, have yielded successful results in ControlNet [67], IP-Adapter [61], T2I-Adapter [31], to name a few. Inspired by recent studies [15, 31, 61], we employ this Parameter-Efficient Fine-Tuning (PEFT) strategy by freezing the parameters of the pre-trained backbone from Trellis [54] and training only the lightweight adapter modules, namely, the *skeleton encoder* and the *injected cross-attention layers*. This SK-Adapter design ensures that the model acquires a robust interpretation of skeletal structure, while fully preserving the powerful generative capabilities of the original foundation model.

Beyond generation, SK-Adapter benefits from our disentangled structural control and has a high potential for flexible 3D editing. With our adapter, our method enables *tuning-free* operations such as addition and replacement of local regions, guided by skeleton prompts. Fig. 1 shows examples generated or edited by SK-Adapter to demonstrate this high potential.

In summary, our main contributions are:

1. We propose **SK-Adapter**, the first framework that achieves skeletal control during native 3D generation. Extensive experiments demonstrate that our method significantly outperforms existing baselines in both structural alignment and generation quality.
2. We construct the **Objaverse-TMS** dataset, comprising 24k high-quality text-mesh-skeleton triplets, addressing the data scarcity bottleneck in structure-guided 3D generation.
3. We demonstrate the capability for flexible **3D editing**. SK-Adapter allows for precise region-specific local editing *based on skeleton prompts*.

2 Related Work

3D generative models. Recent breakthroughs in diffusion models [9, 26, 39], and the increasing availability of large-scale, high-quality 3D datasets [4, 5] have significantly accelerated the evolution of 3D generation. The field has shifted from time-consuming, per-asset optimization [13,35] and multi-view reconstruction-based synthesis [12, 15, 27, 28, 41, 44, 49, 55] toward native 3D generative models [10, 21, 24, 25, 50, 51, 53, 55, 63, 65, 69] like Trellis [54]. These models typi-

cally comprise a variational autoencoder (VAE) [19] and a Diffusion Transformer (DiT) [34] for denoising in latent space and directly operate on 3D structured latents or volumetric representations. These methods unify 3D generation with high fidelity and consistency, eliminating inconsistent multi-view synthesis or inefficient optimization.

Derivatives and extensions of 3D generative models. Following the success of foundational 3D generative frameworks, numerous derivatives have emerged to enhance structural granularity, applicability and editability. To improve generation quality and representational efficiency, works [16,17,22] like Ultra3D [2] have explored advanced architectural designs and diverse 3D priors. To achieve finer control, part-aware methods [1, 7, 42, 58, 59] such as OmniPart [60], and BANG [66] introduce generative mechanisms for part-level decomposition. Meanwhile, the integration of 3D reconstruction with generation, as seen in Amodal3R [52], SAM3D [43] facilitates more robust reconstruction from partial observations. In terms of manipulation, zero-shot editing frameworks [54] such as VoxHammer [23] and Nano3D [62] enable high-fidelity attribute modifications without extensive retraining.

Skeleton-guided generation. Early work AnimatableDreamer [47] generates 3D objects with unstructured skeletons extracted from given videos by canonical score distillation. Many works [11,14,18,31,40,45,46,61,67] in 2D domain have demonstrated that injecting 2D human pose skeletons into diffusion models enables precise manipulation of layout and posture or consistent motion generation for image and video generation. Inspired by this success in 2D domain, a recent work SKDream [56] adopts a "2D lifting" strategy, where arbitrary 3D skeletons are projected into 2D maps to condition a multi-view diffusion model [36], followed by 3D reconstruction [55] and UV refinement to generate 3D assets conditioned on the given skeleton. However, it suffers from spatial ambiguity in 2D skeleton projection and inconsistency across multi-view images during reconstruction. A concurrent work, Posemaster [57], focuses on humanoids and generates 3D humanoids conditioned on a fixed humanoid skeleton with diverse poses.

3 Method

3.1 Problem Formulation

The goal of skeleton-guided 3D generation is to synthesize a high-fidelity 3D asset $\mathcal{X}$ (represented as a 3D latent or mesh) that is consistent with both a textual prompt $\mathcal{T}$ and a specific structural guidance provided by a 3D skeleton $\mathcal{S}$.

Formally, a 3D skeleton $\mathcal{S}$ is defined as a structured tuple $\mathcal{S} = \{J, G\}$, where:

- $J \in \mathbb{R}^{N\times 3}$ denotes the spatial coordinates of N joints in 3D space.
- G represents the topology graph, typically defined by a parent mapping function $p(i)$ or an adjacency matrix $A \in \{0,1\}^{N\times N}$, capturing the kinematic constraints.

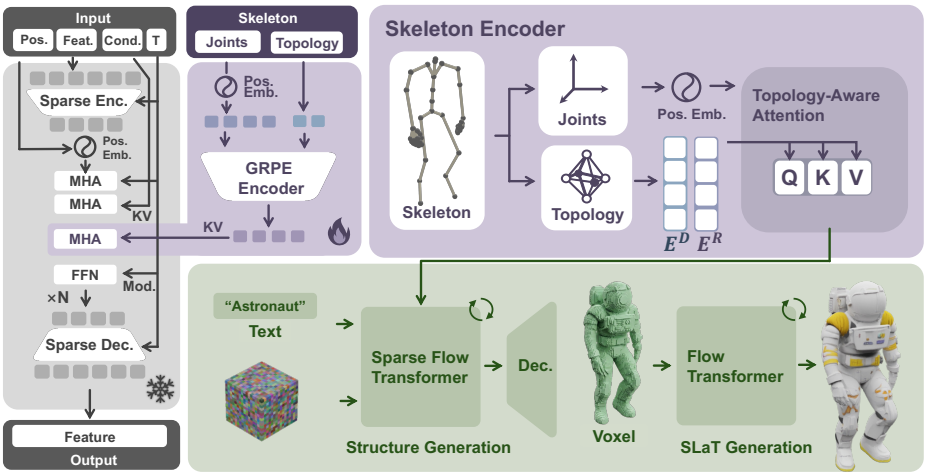

Fig. 2: Overview of the SK-Adapter framework. The GRPE module encodes the 3D skeleton's joints and topology into sparse tokens. These tokens are injected into a frozen pre-trained backbone via trainable cross-attention layers.

The objective is to learn a conditional mapping $\mathcal{F} : (\mathcal{T}, \mathcal{S}) \to \mathcal{X}$. The primary challenge lies in preserving the intricate structural details prescribed by skeletal structure while maintaining the generative quality and diversity underlying the large-scale 3D priors.

3.2 SK-Adapter

We propose SK-Adapter, a 3D skeleton-guided generation framework for precise structural control. As shown in Fig. 2, unlike heavy multi-stage pipelines that rely on ambiguous 2D projections, SK-Adapter utilizes Trellis [54] as backbone and injects joint-based positional tokens into sparse structure transformer blocks. This simple yet effective approach ensures spatial accuracy and high data efficiency, enabling the generation of diverse, precisely controlled 3D assets in seconds.

Topology-Aware Encoding. We represent the input skeleton as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of joint coordinates $J \in \mathbb{R}^{N\times 3}$ and $\mathcal{E}$ represents the hierarchical connectivity G. To capture the intricate spatial and structural relationships within the skeleton, we employ Graph Relative Positional Encoding (GRPE) [33]. We integrate graph properties into the attention maps using two types of node affinity: the topological distance $\mathcal{D}_S$ and relations $\mathcal{R}_S$.

Specifically, for a given pair of joints $i, j \in J$, let q_i and k_j be the query and key representations. We learn distinct embeddings for topological distances $E_q^D, E_k^D \in \mathbb{R}^{(d_{max}+1)\times F}$ and for relations $E_q^R, E_k^R \in \mathbb{R}^{6\times F}$, where F is the latent

feature size. These embeddings are used to form two structural attention biases:

$$a_{ij}^{D} = q_i \cdot E_q^D[D_{ij}] + k_j \cdot E_k^D[D_{ij}], \tag{1}$$

$$a_{ij}^{R} = q_i \cdot E_q^R[R_{ij}] + k_j \cdot E_k^R[R_{ij}], \tag{2}$$

where D_{ij} and R_{ij} denote the topological distance and relation between joints i and j, respectively, and $[\cdot]$ denotes an index in the embedding matrix. These terms are aggregated into the standard attention map, and the final scaled attention score is defined as:

$$a_{ij} = \frac{q_i \cdot k_j + a_{ij}^{D} + a_{ij}^{R}}{\sqrt{F}}. \tag{3}$$

The node features are further enriched by encoding graph information into the values:

$$z_i = \sum_{j=1}^{J} \hat{a}_{ij}(v_j + E_v^D[D_{ij}] + E_v^R[R_{ij}]), \tag{4}$$

where $\hat{a}_{ij}$ is the normalized attention weight. This mechanism yields a skeletal embedding $\mathbf{f}_{skel} \in \mathbb{R}^{J \times F}$ that integrates geometric coordinates with hierarchical topology.

Skeletal Cross Attention Mechanism. To bridge the gap between the 3D skeletal domain and the 3D voxel domain, we introduce a Cross-Attention mechanism. In each block of the flow transformer, the intermediate voxel feature map $\mathbf{h}_{base}$ from the pre-trained backbone is used to query the skeletal information. Consequently, with the standard attention operation:

$$\mathbf{f}_{attn} = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5}$$

where the Query (Q) is derived from the voxel features $\mathbf{h}_{base}$ of the frozen backbone, while the Key (K) and Value (V) are derived from the encoded skeletal features $\mathbf{f}_{skel}$ produced by the GRPE encoder, it allows each spatial voxel to dynamically "attend" to the most relevant skeletal joints, effectively mapping the topological constraints onto the 3D coordinate grid.

To maintain the generative fidelity of the pre-trained model and ensure training stability, the output of the cross-attention layer, $\mathbf{f}_{attn}$, is integrated into the backbone via a non-invasive strategy. We pass $\mathbf{f}_{attn}$ through a zero-initialized linear layer $\mathbf{W}_o$ to obtain $\mathbf{h}_{attn}$. The final hidden state $\mathbf{h}'$ is updated via a residual connection:

$$\mathbf{h}' = \mathbf{h} + \mathbf{h}_{attn} = \mathbf{h} + \mathbf{W}_o\mathbf{f}_{attn}. \tag{6}$$

This design ensures that at the onset of training, SK-Adapter contributes a "null signal," allowing the model to preserve the high-quality generative priors of the original Trellis pipeline. As optimization progresses, the layer gradually learns to modulate the voxel latents according to the skeletal guidance without destabilizing the pre-trained distribution.

Training. The training of SK-Adapter follows the Latent Flow Matching (LFM) [3] paradigm, supervised in the compact latent space of a pre-trained 3D Voxel Autoencoder. For a ground-truth 3D asset, we first obtain its sparse latent representation $\mathbf{z}_0$ using the frozen Voxel Encoder $\mathcal{E}$. We then define a probability path between Gaussian noise $\mathbf{z}_1$ and the target latent $\mathbf{z}_0$. The SK-Adapter-enhanced model v_θ is trained to predict the velocity field that transforms the noise toward the skeletal-conditioned target:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_t} \left\| v_\theta(\mathbf{z}_t, t, \mathbf{c}_{text}, \mathbf{f}_{skel}) - \mathbf{u}_t(\mathbf{z}_0) \right\|^2, \tag{7}$$

where $\mathbf{u}_t(\mathbf{z}_0)$ is the target velocity. During training, the entire transformer backbone remains frozen. Only the GRPE encoder, cross-attention layers, and zero-initialized projection layers are optimized. This prevents catastrophic forgetting of the base model's 3D knowledge while enabling precise structural control.

3.3 Editing

The locality of SLaT allows for region-specific editing by altering voxels and latents in masked areas while leaving other areas intact. To this end, following Trellis [54], we adapt Repaint [29] to our skeleton-conditioned editing. Unlike previous methods that solely rely on text or image prompts, we additionally utilize the skeleton prompt for better structural control.

Given a modified skeleton (e.g., with altered joint angles or topology) and a masked bounding box, we modify the flow matching sampling process to regenerate content strictly within these areas. The generation is conditioned on the unchanged background and the updated skeletal tokens via our SK-Adapter. Consequently, the first stage updates the structural topology to match the new skeletal guidance, and the second stage produces coherent surface details, enabling tuning-free operations such as precise addition and re-posing (replacement of skeleton).

4 Experiments

4.1 Dataset

Training an effective model for skeleton-based 3D generation necessitates synchronized data across three modalities: textual descriptions, 3D meshes, and their corresponding skeletal rigs. However, the absence of such a tripartite dataset in the existing literature poses a significant bottleneck. While recent efforts like SKDream [56] attempt to bridge this gap using autonomous skeleton generation with deterministic algorithms, these automatically generated rigs often suffer from anatomical inconsistencies and lack physical plausibility, which severely limits the model's ability to learn precise structural priors. To facilitate the training of our proposed framework, we curate Objaverse-TMS, a large-scale collection specifically designed for this task. We build Objaverse-TMS from Anymate [6] and CAP3D [30], which are both based on the subsets of Objaverse-v1 [5] and Objaverse-XL [4]. By extracting skeletal structures and captions from

Table 1: Comparison of Objaverse-TMS with other 3D datasets with skeleton annotations.

Dataset	Scale	Skeleton	Text Caption
Rig-XL [64]	14k	Expert	×
Articulation-XL [38]	33k	Expert	×
Articulation-XL2.0 [37]	59k	Expert	×
Anymate [6]	230k	Expert	×
Objaverse-SK [56]	24k	Autonomous	✓
TextuRig [68]	7k	Expert	✓
Objaverse-TMS (Ours)	24k	Expert	✓

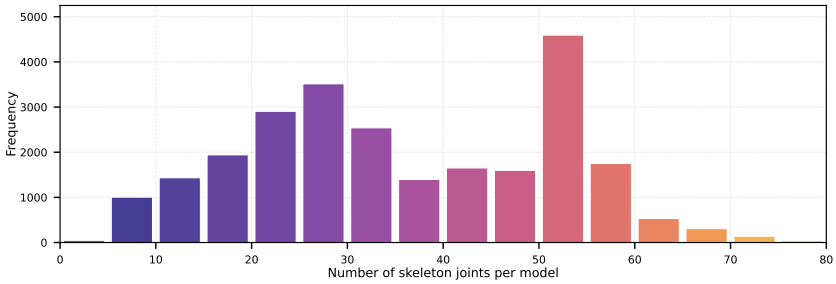

Fig. 3: Distribution of skeleton joint counts in Objaverse-TMS.

Anymate and CAP3D respectively, we establish a high-quality intersection of these modalities. We then process the original raw assets from Objaverse-v1 and Objaverse-XL to derive meshes and normalized voxels with skeletons. After filtering out samples with incomplete rigging, the resulting Objaverse-TMS dataset comprises 24K text-mesh-skeleton triplets, providing a robust foundation for skeleton-conditioned learning. We compare the existing datasets with skeleton annotations in Table 1, report the skeleton joint-count statistics in Fig. 3, and visualize representative dataset samples in Fig. 4.

As shown in Fig. 4, Objaverse-TMS covers a broad range of articulated 3D assets, including humanoids, animals, and other object categories with diverse geometric layouts. The paired skeletons remain spatially aligned with the corresponding meshes, providing explicit joint and bone structures that complement textual descriptions and enable the model to learn fine-grained skeletal control in native 3D space.

4.2 Implementation Details

We train our SK-Adapter on Objaverse-TMS dataset for 200 epochs, with a batch size of 16 and a learning rate of 1×10^{-5}. To support classifier-free guidance, we apply a 10% dropout to the text conditioning, while the skeleton conditioning remains dropout-free.

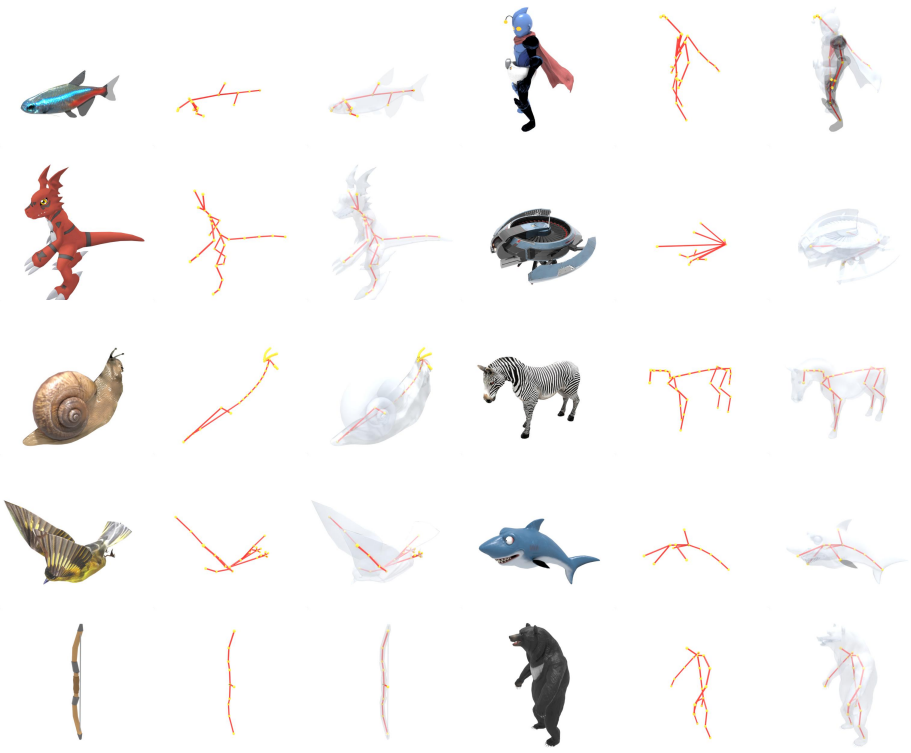

Fig. 4: Representative samples from Objaverse-TMS. Each example shows the rendered 3D asset, its expert-annotated skeleton, and the skeleton-mesh overlay.

4.3 Evaluation Protocol

To ensure a rigorous and unbiased evaluation, we curate a diverse testing set comprising 140 test instances (TMS-eval) sampled from the validation split of Objaverse-TMS. To thoroughly assess the model's generalization capabilities across various topological complexities, this benchmark is strictly balanced to include three primary categories: humanoids, animals, and other objects. Specifically, these assets consist of 54 humanoid figures, 63 animals and 23 objects.

4.4 Evaluation Metrics

We evaluate our method and baselines in terms of structural/textual alignment and visual fidelity. All 2D-based metrics are computed from 12 uniformly distributed renderings of each generated 3D mesh.

Structural and Textual Alignment. To assess conditioning fidelity, we evaluate both structural and textual consistency. For structural alignment, we introduce the ReRigging Score: we apply the rigging model from Anymate [6]

to estimate a skeleton $\hat{S}$ from each generated mesh, then compute its Chamfer Distance to the conditioning skeleton S. Lower scores indicate better structural adherence; applying Anymate to ground-truth meshes gives an oracle reference of 0.2073. For textual alignment, we use CLIP Score to measure the semantic consistency between multi-view renders and text prompts.

Visual Fidelity. To assess the overall visual quality and realism of the generated geometries, we employ PickScore [20] and Kernel Distance (KD-DINO). PickScore evaluates the human-aligned perceptual quality of the generated outputs. To evaluate distributional visual fidelity, we compute KD-DINO by extracting features from the rendered views using the DINOv2 [32] backbone, which robustly measures the global discrepancy between the generated multi-view image distribution and the ground-truth references.

4.5 Baselines Comparison

We benchmark our SK-Adapter, a native but lightweight 3D approach, against representative, state-of-the-art structure-guided generative frameworks. Specifically, we compare against SKDream [56], a multi-view diffusion approach that conditions generation on 2D projected maps of 3D skeletons prior to performing 3D reconstruction. We also modify another native 3D generation method, SpaceControl [8] to our task for comparison. Specifically, SpaceControl is a training-free spatial guidance method built upon the Trellis [54]. SpaceControl injects structural guidance directly into the inference process. To adapt this method to our skeletal conditioning task, we convert the input sparse skeleton into a volumetric mesh representation by modeling joints as spheres and bones as cylinders with the same radius. The process starts from adding noise on conditional skeletons with a time step t_s (set as 0.52 to balance skeleton alignment and generation quality). Then clean 3D assets are generated by denoising steps.

4.6 Quantitative Results

The quantitative comparisons on TMS-eval are summarized in Tables 2 and 3.

Structural and Textual Alignment. Evidenced by Table 2, SK-Adapter achieves state-of-the-art performance in both structural fidelity and semantic consistency. Our method drastically reduces the overall ReRigging Score to 0.2228, significantly outperforming the lifting-based baseline SKDream with 0.2818 and the tuning-free spatial guidance method SpaceControl with 0.2740. This substantial margin confirms that injecting skeletal tokens within the native 3D space effectively mitigates the spatial ambiguity inherent in 2D projection strategies, resulting in assets that faithfully conform to the user-defined topological constraints without overdoing it. Furthermore, SK-Adapter achieves the highest overall CLIP Score, demonstrating that our lightweight adapter maintains robust text-to-3D cross-modal alignment.

Visual Fidelity. Beyond structural precision, our framework maintains superior visual quality, as detailed in Table 3. SK-Adapter achieves the highest

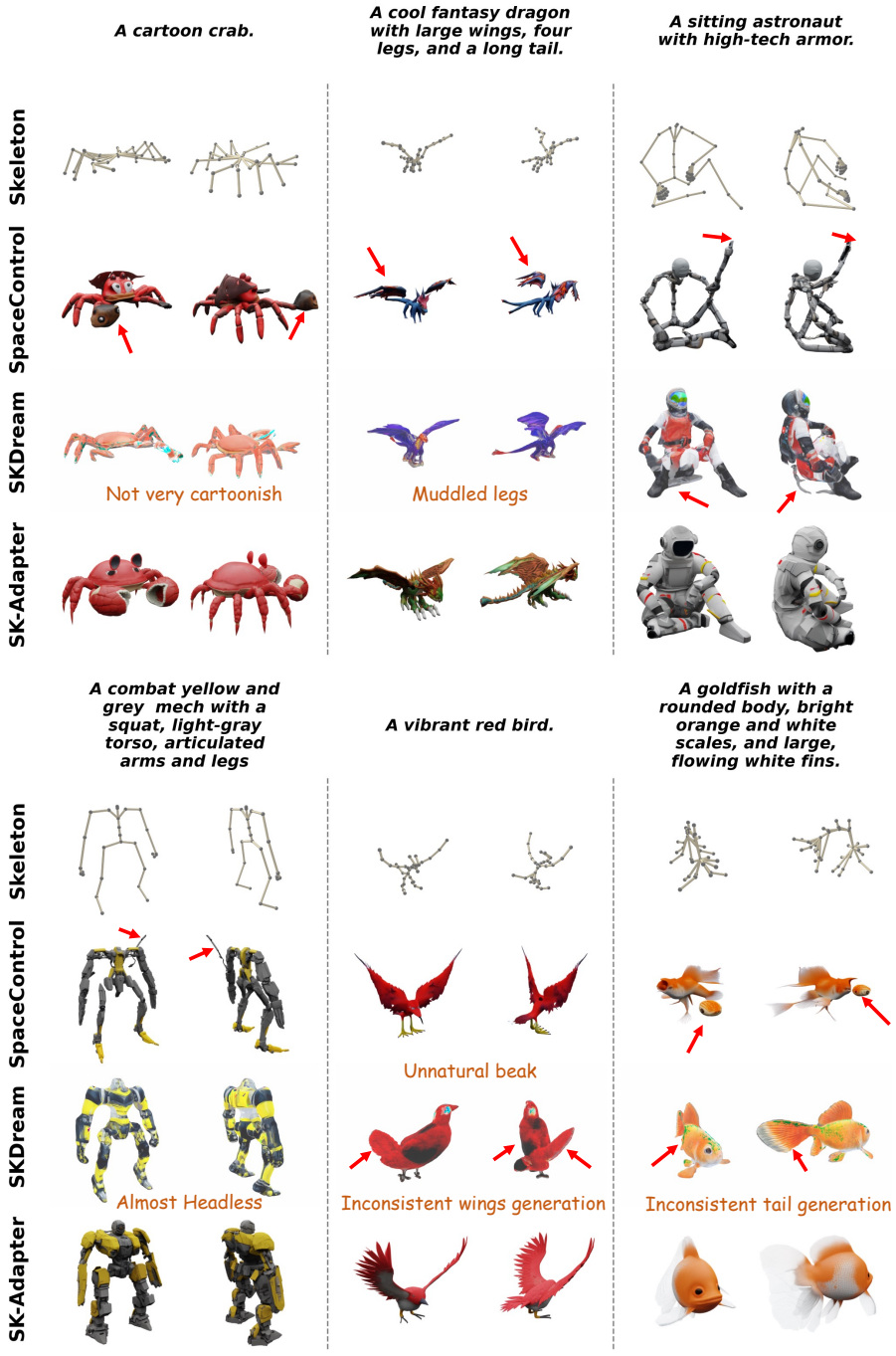

Fig. 5: Qualitative comparison of our SK-Adapter with the baselines. For the baseline generations, arrows indicate structural inconsistency with the given skeleton. Short captions indicate one of the apparent problems in their generated results.

Table 2: Quantitative Comparison on TMS-eval (Structural and Textual Alignment). We evaluate the methods based on Structural Alignment (ReRigging Score) and Textual Alignment (CLIP Score) across overall, humanoid, animal, and other categories. ↓ indicates lower is better, ↑ indicates higher is better. **Bold** indicates the best performance.

	ReRigging Score ↓				**CLIP Score ↑**			
Method	Overall	Humanoid	Animal	Other	Overall	Humanoid	Animal	Other
SKDream [56]	0.2818	0.2385	0.2730	0.4075	25.65	25.33	26.34	**24.52**
SpaceControl [8]	0.2740	0.2282	0.2898	0.3386	25.66	25.64	26.50	23.43
SK-Adapter (Ours)	**0.2228**	**0.1740**	**0.2415**	**0.2861**	**26.16**	**26.28**	**27.05**	23.43

Table 3: Quantitative Comparison on TMS-eval (Visual Fidelity). We evaluate the methods based on Visual Fidelity (PickScore, KD-DINO) across overall, humanoid, animal, and other categories. ↓ indicates lower is better, ↑ indicates higher is better. **Bold** indicates the best performance.

	PickScore ↑				**KD-DINO ↓**			
Method	Overall	Humanoid	Animal	Other	Overall	Humanoid	Animal	Other
SKDream [56]	20.46	20.01	20.88	20.40	1.3809	2.1106	1.7809	3.2435
SpaceControl [8]	20.55	20.18	20.90	**20.45**	1.7821	2.1719	2.5034	3.6304
SK-Adapter (Ours)	**21.01**	**20.74**	**21.45**	20.44	**0.7778**	**1.1998**	**1.2673**	**2.3765**

overall PickScore of 21.01, indicating a strong human preference for our generated geometries. Also, our method significantly lowers the KD-DINO metric to 0.7778, compared to 1.3809 for SKDream and 1.7821 for SpaceControl. These results strongly indicate that our strategy of freezing the backbone effectively preserves the powerful generative priors of the foundation model, synthesizing high-frequency details and coherent textures without suffering from the geometric artifacts or texture degradation typically observed in multi-stage reconstruction pipelines.

Running Time. Under the same hardware conditions, our SK-Adapter as well as SpaceControl can generate a 3D asset in less than 15 seconds while SKDream requires around 40 seconds to generate one sample on average.

4.7 Qualitative Results

Visual comparisons are presented in Fig. 5. Our method consistently outperforms baseline approaches by demonstrating significantly better skeleton interpretation, finer structural alignment, and stricter adherence to text prompts. Furthermore, it yields higher visual fidelity and more detailed geometries. While our approach excels at generating coherent and intricate structures, alternative methods suffer from notable quality degradation. SKDream exhibits structural distortions and vague or missing texture details caused by the multi-view inconsistencies inherent in its underlying 2D generative models; SpaceControl

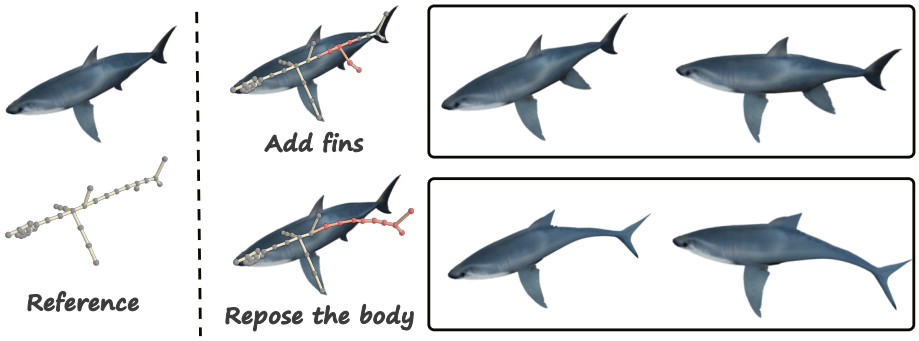

Fig. 6: SK-Adapter enables flexible editing with skeleton, including addition and re-posing.

produces assets that are overly constrained by the skeleton, often leading to fragmented or broken topologies.

4.8 Application

Editing. Fig. 6 illustrates sample editing sequences of 3D assets, involving addition and re-posing (replacement). Given an edited skeleton with the identified bounding box, our method enables skeleton-based flexible editing on existing 3D assets, where the generated mesh of the locally-edited skeleton blends seamlessly with the original mesh, preserving the underlying structural and textural coherence.

Animation. Our skeleton-conditioned generated 3D assets can be seamlessly integrated into standard animation pipelines. Our approach outputs static 3D assets that can be rigged and driven by skeletal motions. Specifically, we use the off-the-shelf skinning model from Anymate [6] to compute skinning weights using the input skeleton and our generated 3D assets. Given the skinning weights, we apply Linear Blend Skinning (LBS) to deform the asset according to the input skeleton sequences.

4.9 Ablation Study

To validate the effectiveness of our key architectural components, we conduct an ablation study by evaluating different variants of SK-Adapter. Each variant is trained on an 8k subset of Objaverse-TMS for 200 epochs. Specifically, we implement the ablation study on the following components:

Skeletal Cross Attention. In this variant, we remove the dedicated skeletal cross-attention layers. Instead, we directly concatenate the encoded skeleton features with the text features, thus forcing both modalities to share a single

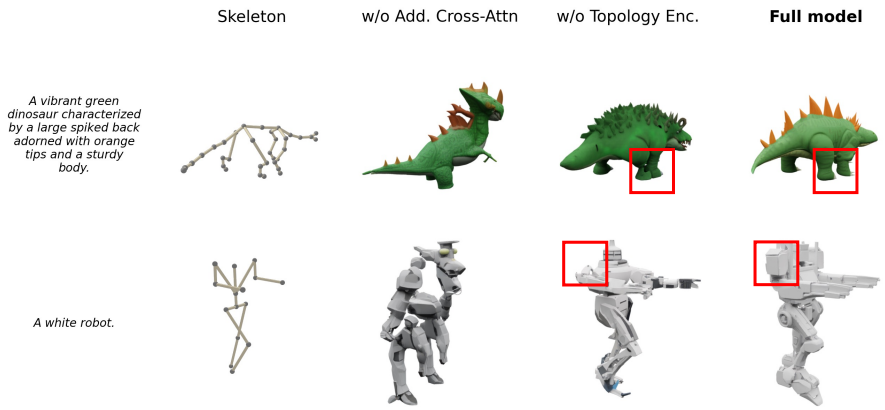

Fig. 7: Comparison of different architecture designs. Without Cross-Attention, the model tends to collapse given complex skeleton conditions. In contrast, the topology-aware encoder helps the model better capture skeleton structures, as indicated by the bounding boxes.

cross-attention layer. In doing so, we evaluate the necessity of a dedicated cross-attention mechanism for injecting structural guidance without entangling it with textual semantics.

Topology-Aware Encoding. In this variant configuration, the topological graph input and GRPE encoder are removed. Consequently, the network is conditioned solely on the positional embeddings of the spatial joint coordinates. This tests the importance of explicitly modeling hierarchical connectivity and topological constraints for accurate 3D articulation.

Tables 4 and 5 tabulate the quantitative results tested on TMS-eval. As shown in Table 4, the Full Model achieves the most robust structural control, yielding the best ReRigging Score of 0.2355 overall. Removing the dedicated skeletal cross-attention causes a severe degradation in structural alignment, effectively doubling the ReRigging Score to 0.5049. Similarly, discarding the topological encoding negatively impacts the model's ability to accurately articulate the 3D generation.

Crucially, Tables 4 and 5 demonstrate that the Full Model's significant improvements in structural control do not come at the expense of generation quality. Although the variant without topology encoding shows very marginal gains in CLIP Score and KD-DINO, the Full Model achieves the highest PickScore across all categories. The Full Model therefore offers the optimal architectural balance between fine-grained skeleton controllability and high-quality 3D generation.

Qualitative comparisons are shown in Fig. 7. Without Cross-Attention, the model tends to collapse given complex skeleton conditions. Although the model without a topological encoder can generate 3D assets that follow the skeleton joints, the topology-aware encoder helps it better capture skeleton structures.

Table 4: Ablation Study (Structural and Textual Alignment). Impact of individual architectural components on structural control and textual alignment across overall, humanoid, and animal categories. ↓ indicates lower is better, ↑ indicates higher is better. **Bold** indicates the best performance, underline indicates the second best.

	ReRigging Score ↓			CLIP Score ↑		
Variant	Overall	Humanoid	Animal	Overall	Humanoid	Animal
w/o Add. Cross-Attn	0.5049	0.4230	0.5353	24.71	25.16	25.27
w/o Topology Enc.	0.2527	0.1885	0.2747	**26.14**	**26.42**	26.87
Full Model	**0.2355**	**0.1832**	**0.2531**	26.11	26.16	**26.99**

Table 5: Ablation Study (Visual Fidelity). Impact of individual architectural components on visual fidelity across overall, humanoid, and animal categories. ↓ indicates lower is better, ↑ indicates higher is better. **Bold** indicates the best performance, underline indicates the second best.

	PickScore ↑			KD-DINO ↓		
Variant	Overall	Humanoid	Animal	Overall	Humanoid	Animal
w/o Add. Cross-Attn	20.54	20.39	20.86	1.0914	1.2346	1.9521
w/o Topology Enc.	20.90	20.61	21.29	**0.7865**	**1.1539**	**1.3378**
Full Model	**20.94**	**20.61**	**21.36**	0.8616	1.1655	1.3812

5 Conclusion

In this paper, we propose SK-Adapter, a lightweight framework that achieves precise skeleton-guided structural control for native 3D generation. By encoding 3D skeletons as sparse, topology-aware spatial tokens and seamlessly injecting them into a frozen 3D flow transformer, our method ensures faithful structural alignment while fully preserving the powerful generative priors of the foundation model. These capabilities directly facilitate the creation of production-ready, articulable 3D assets. Consequently, by treating skeletons as control signals, SK-Adapter bridges the gap between abstract semantic synthesis and explicit topological articulation, paving the way toward more interpretable and controllable 3D generation.

References

1. Chen, M., Wang, J., Shapovalov, R., Monnier, T., Jung, H., Wang, D., Ranjan, R., Laina, I., Vedaldi, A.: Autopartgen: Autogressive 3d part generation and discovery (2025), https://arxiv.org/abs/2507.13346
2. Chen, Y., Li, Z., Wang, Y., Zhang, H., Li, Q., Zhang, C., Lin, G.: Ultra3d: Efficient and high-fidelity 3d generation with part attention (2025), https://arxiv.org/abs/2507.17745
3. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space (2023), https://arxiv.org/abs/2307.08698

4. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects (2023), https://arxiv.org/abs/2307.05663
5. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects (2022), https://arxiv.org/abs/2212.08051
6. Deng, Y., Zhang, Y., Geng, C., Wu, S., Wu, J.: Anymate: A dataset and baselines for learning 3d object rigging (2025), https://arxiv.org/abs/2505.06227
7. Dong, S., Ding, L., Chen, X., Li, Y., Wang, Y., Wang, Y., Wang, Q., Kim, J., Gao, C., Huang, Z., Wang, Z., Xue, T., Xu, D.: From one to more: Contextual part latents for 3d generation (2025), https://arxiv.org/abs/2507.08772
8. Fedele, E., Engelmann, F., Huang, I., Litany, O., Pollefeys, M., Guibas, L.: Spacecontrol: Introducing test-time spatial control to 3d generative modeling (2025), https://arxiv.org/abs/2512.05343
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020), https://arxiv.org/abs/2006.11239
10. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d (2024), https://arxiv.org/abs/2311.04400
11. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation (2024), https://arxiv.org/abs/2311.17117
12. Huang, X., Cheung, K.C., Cong, R., See, S., Wan, R.: Stereo-gs: Multi-view stereo vision model for generalizable 3d gaussian splatting reconstruction (2026), https://arxiv.org/abs/2507.14921
13. Huang, Y., Wang, J., Zeng, A., Cao, H., Qi, X., Shi, Y., Zha, Z.J., Zhang, L.: Dreamwaltz: Make a scene with complex 3d animatable avatars (2023), https://arxiv.org/abs/2305.12529
14. Huang, Y., Wang, J., Zeng, A., Zha, Z.J., Zhang, L., Liu, X.: Dreamwaltz-g: Expressive 3d gaussian avatars from skeleton-guided 2d diffusion (2024), https://arxiv.org/abs/2409.17145
15. Huang, Z., Guo, Y.C., Wang, H., Yi, R., Ma, L., Cao, Y.P., Sheng, L.: Mv-adapter: Multi-view consistent image generation made easy (2024), https://arxiv.org/abs/2412.03632
16. Jia, T., Yan, D., Hao, D., Li, Y., Zhang, K., He, X., Li, L., Wang, Y., Chen, J., Jiang, L., Yin, Q., Quan, L., Chen, Y.C., Yuan, L.: Ultrashape 1.0: High-fidelity 3d shape generation via scalable geometric refinement (2025), https://arxiv.org/abs/2512.21185
17. Jin, B., Li, W., Gu, S., Zheng, Y., Zheng, Y., Zhou, Z., Yao, Y.: Uniart: Unified 3d representation for generating 3d articulated objects with open-set articulation (2025), https://arxiv.org/abs/2511.21887
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), https://arxiv.org/abs/2304.04269
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022), https://arxiv.org/abs/1312.6114
20. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=G5RwHpBUv0

21. Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., Yang, X., Yang, M., Yang, S., Feng, Y., Zhang, S., Huang, X., Luo, D., Yang, F., Yang, F., Wang, L., Liu, S., Tang, Y., Cai, Y., He, Z., Liu, T., Liu, Y., Jiang, J., Linus, Huang, J., Guo, C.: Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details (2025), https://arxiv.org/abs/2506.16504
22. Lai, Z., Zhao, Y., Zhao, Z., Liu, H., Lin, Q., Huang, J., Guo, C., Yue, X.: Lattice: Democratize high-fidelity 3d generation at scale (2025), https://arxiv.org/abs/2512.03052
23. Li, L., Huang, Z., Feng, H., Zhuang, G., Chen, R., Guo, C., Sheng, L.: Voxhammer: Training-free precise and coherent 3d editing in native 3d space (2025), https://arxiv.org/abs/2508.19247
24. Li, W., Liu, J., Yan, H., Chen, R., Liang, Y., Chen, X., Tan, P., Long, X.: Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner (2025), https://arxiv.org/abs/2405.14979
25. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., Cao, Y.P.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models (2025), https://arxiv.org/abs/2502.06608
26. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
27. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image (2024), https://arxiv.org/abs/2309.03453
28. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., Wang, W.: Wonder3d: Single image to 3d using cross-domain diffusion (2023), https://arxiv.org/abs/2310.15008
29. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models (2022), https://arxiv.org/abs/2201.09865
30. Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3d captioning with pretrained models (2023), https://arxiv.org/abs/2306.07279
31. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models (2023), https://arxiv.org/abs/2302.08453
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
33. Park, W., Chang, W., Lee, D., Kim, J., won Hwang, S.: Grpe: Relative positional encoding for graph transformer (2022), https://arxiv.org/abs/2201.12787
34. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), https://arxiv.org/abs/2212.09748
35. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion (2022), https://arxiv.org/abs/2209.14988
36. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation (2024), https://arxiv.org/abs/2308.16512
37. Song, C., Li, X., Yang, F., Xu, Z., Wei, J., Liu, F., Feng, J., Lin, G., Zhang, J.: Puppeteer: Rig and animate your 3d models (2025), https://arxiv.org/abs/2508.10898

38. Song, C., Zhang, J., Li, X., Yang, F., Chen, Y., Xu, Z., Liew, J.H., Guo, X., Liu, F., Feng, J., Lin, G.: Magicarticulate: Make your 3d models articulation-ready (2025), https://arxiv.org/abs/2502.12135
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022), https://arxiv.org/abs/2010.02502
40. Tan, S., Gong, B., Wang, X., Zhang, S., Zheng, D., Zheng, R., Zheng, K., Chen, J., Yang, M.: Animate-x: Universal character image animation with enhanced motion representation (2024), https://arxiv.org/abs/2410.10306
41. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation (2024), https://arxiv.org/abs/2402.05054
42. Tang, J., Lu, R., Li, Z., Hao, Z., Li, X., Wei, F., Song, S., Zeng, G., Liu, M.Y., Lin, T.Y.: Efficient part-level 3d object generation via dual volume packing (2025), https://arxiv.org/abs/2506.09980
43. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025), https://arxiv.org/abs/2511.16624
44. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion (2024), https://arxiv.org/abs/2403.12008
45. Wang, R., Hu, T., Huang, K., Su, Z., Yi, R., Ma, L.: Poseanything: Universal pose-guided video generation with part-aware temporal coherence (2025), https://arxiv.org/abs/2512.13465
46. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for realistic human dance generation (2024), https://arxiv.org/abs/2307.00040
47. Wang, X., Wang, Y., Ye, J., Wang, Z., Sun, F., Liu, P., Wang, L., Sun, K., Wang, X., He, B.: Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation (2024), https://arxiv.org/abs/2312.03795
48. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation (2023), https://arxiv.org/abs/2305.16213
49. Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: Crm: Single image to 3d textured mesh with convolutional reconstruction model (2024), https://arxiv.org/abs/2403.05034
50. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer (2024), https://arxiv.org/abs/2405.14832
51. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Yang, Y., Bao, Y., Qian, J., Zhu, S., Cao, X., Torr, P., Yao, Y.: Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention (2025), https://arxiv.org/abs/2505.17412
52. Wu, T., Zheng, C., Guan, F., Vedaldi, A., Cham, T.J.: Amodal3r: Amodal 3d reconstruction from occluded 2d images (2025), https://arxiv.org/abs/2503.13439
53. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J.: Native and compact structured latents for 3d generation (2025), https://arxiv.org/abs/2512.14692

54. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation (2025), https://arxiv.org/abs/2412.01506
55. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models (2024), https://arxiv.org/abs/2404.07191
56. Xu, Y., Yang, Z., Yang, Y.: Skdream: Controllable multi-view and 3d generation with arbitrary skeletons. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 314–325 (June 2025)
57. Yan, H., Luo, K., Li, W., Zhang, K., Liang, Y., Huang, J., Guo, C., Tan, P.: Posemaster: A unified 3d native framework for stylized pose generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 34292–34302 (June 2026)
58. Yan, X., Xu, J., Li, Y., Ma, C., Yang, Y., Wang, C., Zhao, Z., Lai, Z., Zhao, Y., Chen, Z., Guo, C.: X-part: high fidelity and structure coherent shape decomposition (2025), https://arxiv.org/abs/2509.08643
59. Yang, Y., Guo, Y.C., Huang, Y., Zou, Z.X., Yu, Z., Li, Y., Cao, Y.P., Liu, X.: Holopart: Generative 3d part amodal segmentation (2025), https://arxiv.org/abs/2504.07943
60. Yang, Y., Zhou, Y., Guo, Y.C., Zou, Z.X., Huang, Y., Liu, Y.T., Xu, H., Liang, D., Cao, Y.P., Liu, X.: Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion (2025), https://arxiv.org/abs/2507.06165
61. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023), https://arxiv.org/abs/2308.06721
62. Ye, J., Xie, S., Zhao, R., Wang, Z., Yan, H., Zu, W., Ma, L., Zhu, J.: Nano3d: A training-free approach for efficient 3d editing without masks (2025), https://arxiv.org/abs/2510.15019
63. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models (2023), https://arxiv.org/abs/2301.11445
64. Zhang, J.P., Pu, C.F., Guo, M.H., Cao, Y.P., Hu, S.M.: One model to rig them all: Diverse skeleton rigging with unirig (2025), https://arxiv.org/abs/2504.12451
65. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets (2024), https://arxiv.org/abs/2406.13897
66. Zhang, L., Zhang, Q., Jiang, H., Bai, Y., Yang, W., Xu, L., Yu, J.: Bang: Dividing 3d assets via generative exploded dynamics. ACM Transactions on Graphics **44**(4), 1–21 (Jul 2025). https://doi.org/10.1145/3730840, http://dx.doi.org/10.1145/3730840
67. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023), https://arxiv.org/abs/2302.05543
68. Zhao, R., Zheng, H., Yang, Z., Fan, H., Yang, Y.: Stroke3d: Lifting 2d strokes into rigged 3d model via latent diffusion models (2026), https://arxiv.org/abs/2602.09713
69. Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., Fu, B., Chen, T., Yu, G., Gao, S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation (2023), https://arxiv.org/abs/2306.17115