YouDream : Generating Anatomically Controllable Consistent Text-to-3D Animals

Sandeep Mishra[†*], **Oindrila Saha**[‡*], **and Alan C. Bovik**[†]
[†]University of Texas at Austin, [‡]University of Massachusetts Amherst
sandy.mishra@utexas.edu, osaha@umass.edu, bovik@ece.utexas.edu

Abstract

3D generation guided by text-to-image diffusion models enables the creation of visually compelling assets. However previous methods explore generation based on image or text. The boundaries of creativity are limited by what can be expressed through words or the images that can be sourced. We present YouDream, a method to generate high-quality anatomically controllable animals. YouDream is guided using a text-to-image diffusion model controlled by 2D views of a 3D pose prior. Our method is capable of generating novel imaginary animals that previous text-to-3D generative methods are unable to create. Additionally, our method can preserve anatomic consistency in the generated animals, an area where prior approaches often struggle. Moreover, we design a fully automated pipeline for generating commonly observed animals. To circumvent the need for human intervention to create a 3D pose, we propose a multi-agent LLM that adapts poses from a limited library of animal 3D poses to represent the desired animal. A user study conducted on the outcomes of YouDream demonstrates the preference of the animal models generated by our method over others. Visualizations and code are available at https://youdream3d.github.io/.

1 Introduction

Text-to-3D generative modeling using diffusion models has seen fast-paced growth recently with methods utilizing text-to-image (T2I) Poole et al. (2022); Chen et al. (2023a); Zhu et al. (2023); Seo et al. (2023), (text+camera)-to-image (TC2I) Shi et al. (2023); Li et al. (2023) and (image+camera)-to-image (IC2I) Liu et al. (2023); Wang and Shi (2023); Ye et al. (2023) diffusion models. These methods are widely accepted by AI enthusiasts, content creators, and 3D artists to create high-quality 3D content. However, generating 3D assets using such methods is dependent on what can be expressed through text or the availability of an image faithful to the user's imagination. In this work, we provide more control to the artist to bring their creative imagination to life. YouDream can generate high-quality 3D animals based on any 3D skeleton, by utilizing a 2D pose-controlled diffusion model which generates images adhering to 2D views of a 3D pose. Using depth, edge, and scribble has also been explored for controllable image generation Zhang et al. (2023). However, in a 3D context, pose offers both 3D consistency as well as room for creativity. Other controls are restrictive as edge/depth/boundary of 2D views of a pre-existing object is used to provide control, thus limiting the generated shape to be very similar to the existing asset. We show that the multi-view consistency offered by our 3D pose prior results in the generation of anatomically and geometrically consistent animals. Creating 3D pose control also requires minimal human effort. To further alleviate this effort, we also present a multi-agent LLM setup that generates 3D poses for novel animals commonly observed in nature.

[*]Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

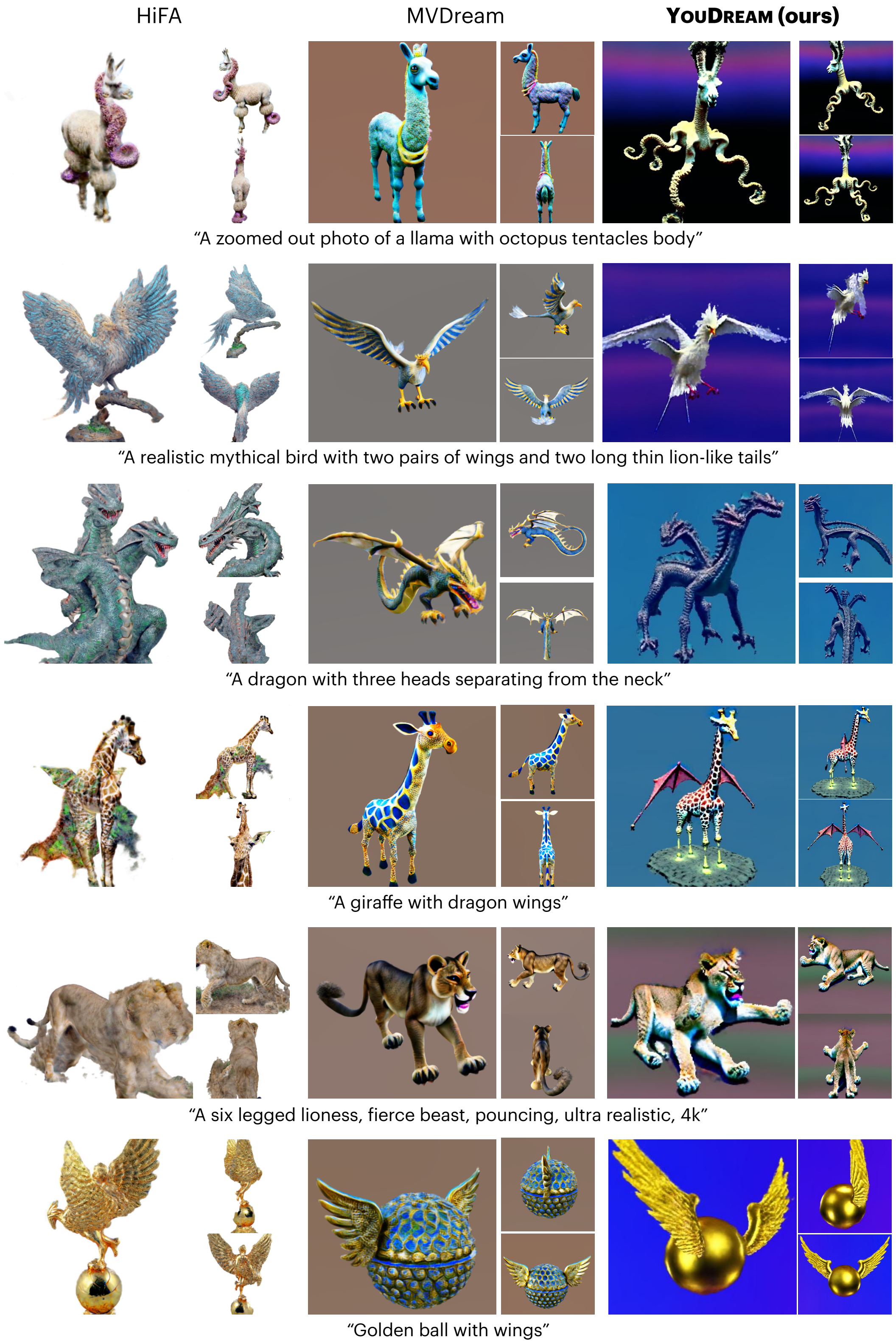

Figure 1: **Creating unreal creatures.** Our method generates imaginary creatures based on an artist's creative control. We show that these creatures cannot be generated faithfully only based on text. Each row depicts a 3D animal generated by HiFA, MVDream, and YOUDREAM (left to right) using the prompt mentioned below the row. We present 3D pose controls used to create these in the Sec. F (results best viewed zoomed in).

3D generation guided by T2I models involves using the gradient computed by Score Distillation Sampling (SDS) Poole et al. (2022) to optimize a 3D representation such as a NeRF Mildenhall et al. (2021). During any intermediate step of the training, a rendered image captured by a random camera is added to Gaussian noise and passed to a T2I diffusion model, along with a directional prompt. The diffusion model estimates the added noise, which in turn is used to create a denoised image. In effect, this process pushes the rendered image of the NeRF representation slightly closer to the denoised image during each iteration. Thus, any unwanted semantic or perceptual issues arising in the denoised image are also transferred to the NeRF. This is especially problematic for deformable objects such as animals, where variations in pose over views often results in the Janus-head problem, dehydrated assets, and geometric and anatomical inconsistencies.

TC2I diffusion models, which encode camera parameters and train using 3D objects from various views learn multi-view consistency and thus are able to produce better geometries. However, they lack in diversity owing to the limited variation in training data, as compared to text-to-image models. Along with this, methods using IC2I diffusion models also face the problems arising from Novel View Synthesis (NVS), which requires hallucination of unseen regions along with accurate geometric transformation of observed parts. While these camera guided diffusion models perform better than T2I models in many cases, their limited diversity and lack of control limit the creativity of their users. By utilizing a 3D pose prior, YOUDREAM consistently outperforms previous methods that use T2I diffusion models, in terms of generating biologically plausible animals. Despite not being trained on any 3D data, our method also outperforms the 3D-aware TC2I diffusion model MVDream Shi et al. (2023) in text-to-3D animal generation in terms of "Naturalness", "Text-Image Alignment" and CLIP score (see Sec. 4).

3D-consistency for human avatar creation has been explored extensively in recent works Cao et al. (2023); Huang et al. (2024); Kolotouros et al. (2024); Hong et al. (2022); Zhang et al. (2024a, 2022). These models rely on a 3D human pose and shape prior, usually the SMPL Loper et al. (2023) or SMPL-X Pavlakos et al. (2019) model. This strategy can represent a variety of geometrically consistent human avatars. However, representing the animal kingdom is challenging owing to its immense diversity which cannot be represented using any existing parametric models. Sizes and shapes vary considerably across birds, reptiles, mammals, and amphibians, hence until now, no single shape or pose prior exists that can represent all tetrapods. Parametric models such as SMAL Zuffi et al. (2017) and MagicPony Wu et al. (2023b) suffer from severe diversity issues, and hence cannot be used as pose or shape prior. Thus to circumvent human effort in generating a 3D pose prior for animals prevalent in nature, we present a method for automatic generation of diverse 3D poses using a multi-agent LLM supported by a small library of animal 3D poses. Additionally, we present a method to automatically generate an initial shape based on a 3D pose, which is utilized for NeRF initialization.

In summary, YOUDREAM offers the following key contributions:

- a TetraPose ControlNet, trained on tetrapod animals across various families, that enables the generation of diverse animals at test time, both real and unreal.
- a multi-agent LLM that can generate the 3D pose of any desired animal in a described state, supported by a small library of 16 predefined animal 3D poses for reference.
- a user-friendly tool to create/modify 3D poses for unreal creatures. The same tool automatically generates an initial shape based on the 3D skeleton.
- a pipeline to generate geometrically and anatomically consistent animals based on an input text by adhering to a 3D pose prior.

2 Related Work

The field of **3D animal generation** has rapidly advanced due to studies that offer methods and insights for modeling animal structures and movements in 3D. SMAL Zuffi et al. (2017) introduced a method to fit a parametric 3D shape model, derived from 3D scans, to animal images using 2D keypoints and segmentation masks, with extensions to multi-view images Zuffi et al. (2018). The variety of animals able to be represented by SMAL is severely limited. Subsequent efforts, such as LASSIE Yao et al. (2022, 2023, 2024), have focused on deriving 3D shapes directly from smaller image collections by identifying self-supervised semantic correspondences to discover 3D parts. Succeeding work represent animals using a parametric model Jakab et al. (2023); Wu et al. (2023a,b); Li et al. (2024) learnt from images or videos. Despite these advances, these methods are class-specific and lack in the diversity of animals that can be represented. YOUDREAMis able to generate a great variety of animals including those that have not been observed previously with higher details (Fig. 16).

High quality **text-to-3D asset generation** has been fueled by the availability of large-scale diverse datasets of text-image pairs and the success of text-to-image contrastive and generative models trained on them. Contrastive methods such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021) learn a common embedding between visual and natural language domains. Generative methods like Imagen Saharia et al. (2022) and Stable Diffusion Rombach et al. (2022) utilize a diffusion model to learn to generate images given text latents. These methods inherently learn to understand the appearance of entities across various views and poses. Text-to-3D generative modeling methods Mohammad Khalid et al. (2022); Jain et al. (2022); Wang et al. (2023); Poole et al. (2022) exploit this information by using these text-image models to guide the creation of 3D representations by NeRFs Mildenhall et al. (2021). The quality of 3D assets produced by these early methods suffer from several issues such as smooth geometries, saturated appearances, as well as geometric issues such as the Janus (multi-head) problem. Subsequent recent methods have ameliorated these problems by the use of modified loss functions Wang et al. (2024); Zhu et al. (2023), using Deep Marching Tetrahedra Shen et al. (2021) for 3D representation Chen et al. (2023a), and modified negative prompt weighing strategies Armandpour et al. (2023). However these methods still fail to produce anatomically correct animals, often producing implausible geometries or even extra or insufficient limbs. Prior work 3DFuse Seo et al. (2023) uses sparse point clouds predicted from images as depth control for T2I diffusion, however still produces anatomically inconsistent animals due to the inaccuracy of image-to-point cloud predictors and a high dependency on the initial generated image (Fig. 4). Recently, 3D-aware diffusion models trained on paired text-3D datasets by encoding camera parameters have been used to generate 3D assets Tang et al. (2023); Shi et al. (2023). As these methods learn using various views of 3D objects, they rarely produce geometric inconsistency. However these methods are limited by the variety of 3D data available, which is quite scarce as compared to image data that T2I diffusion models have been trained on. These are trained using 3D object databases such as Objaverse Deitke et al. (2023) and Objaverse-XL Deitke et al. (2024) which are considerably smaller than text-image paired datasets such as LAION-5B Schuhmann et al. (2022) used for training T2I diffusion models. Thus, they often struggle to follow the text input faithfully in case of complex prompts (Fig. 1). By comparison, our method accurately follows the text prompt owing to the use of T2I diffusion models trained on vast image data. YOUDREAM strictly adheres to input 3D pose priors, thus producing geometrically consistent and anatomically correct animals.

Large Language Models (LLMs) have been explored in the context of **3D generation and editing** previously. LLMs have been used Yin et al. (2023); Siddiqui et al. (2023) to generate and edit shapes using an embedding space trained on datasets such as ShapeNet Chang et al. (2015). Prior work have also used LLMs to generate code for 3D modeling software, such as Blender, to create objects Yuan et al. (2024) and scenes Sun et al. (2023); Hu et al. (2024). These methods produce impressive results suggesting at LLMs' 3D understanding capability, but explore limited variety of generation often limited to shapes, or generate layouts/scenes. 3D pose generation with LLMs using text as input has been recently explored for humans. ChatPose Feng et al. (2024) and MotionGPT Zhang et al. (2024b) generate pose parameters for a SMPL model based on textual input. LLMs have also been previously shown to accurately reason about anatomical differences of animals Menon and Vondrick (2022); Saha et al. (2024). In this work, we show a novel application of off-the-shelf LLMs for generalized 3D pose generation based on the name of an animal supported by a library of animal 3D poses.

User-controlled generation has been introduced in several studies Zhang et al. (2023); Mou et al. (2024), and has gained widespread adoption among artists for crafting remarkable illustrations, including artistic QR codes to interior designs. However, the use of user control in 3D is still under-explored. Recent works such as MVControl Li et al. (2023) and Control3D Chen et al. (2023b) guide the 3D generation process using a 2D condition image of a single view. By contrast, the generation process in YOUDREAM is guided using 2D views of a 3D pose which is dependent on the sampled camera pose. This strategy not only allows YOUDREAM to take in specialized user control but also ensures multi-view geometric consistency.

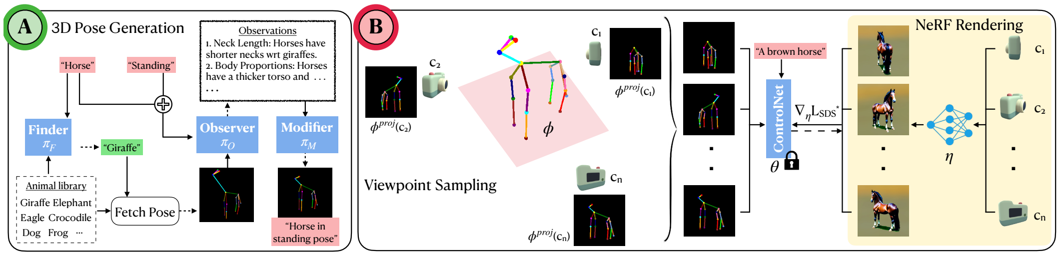

Figure 2: **Automatic pipeline for 3D animal generation.** Given the name of an animal and textual pose description, we utilize a multi-agent LLM to generate a 3D pose (ϕ) supported by a small library of animal names paired with 3D poses. With the obtained 3D pose, we train a NeRF to generate the 3D animal guided by a diffusion model controlled by 2D views (ϕ^{proj}) of ϕ.

3 Method

Multi-view sampling from T2I diffusion models for 3D generation is guided using directional prompt extensions such as "<user_text>, front view" and "<user_text>, side view". Such a control signal is ambiguous due to, 1) directional text remaining unchanged over a range of camera parameters, 2) T2I diffusion models generating deformable entities in various poses for the same view. Thus we utilize 3D pose as a stronger guidance to maintain consistency over different views. To do this in a 3D consistent manner we design 1) a model to generate 2D image samples following the projection ϕ^{proj} of a 3D pose ϕ of an animal, 2) a method to generate the 3D pose ϕ of a novel animal y using a limited library of 3D poses (Φ) of commonly observed animals in nature and a multi-agent LLM pose editor, and 3) a method to create 3D animals given an animal name y and 3D pose ϕ. Our 3D model is represented using Neural Radiance Fields (NeRF Mildenhall et al. (2021)).

3.1 TetraPose ControlNet

To train a model to follow pose control we require images of animals with annotated poses. Datasets released by Banik et al. (2021) and Ng et al. (2022) provide 2D pose annotations of animal images spanning a large number of species, compared to the limited diversity available in 3D animal pose datasets Xu et al. (2023); Badger et al. (2020). We thus utilize these 2D pose datasets by learning to map the 2D pose of an animal to its captured image. We define such datasets of animal species y_j, corresponding animal images x_j, and their 2D pose ϕ_j^{proj} as the set $\mathcal{D} = \{(x_j, \phi_j^{proj}, y_j)\}_{j=1}^{J}$, where $J = |\mathcal{D}|$ is the number of image-pose pairs in the dataset. This learned mapping can then be used to generate multi-view image samples consistent with a 3D pose ϕ. The mapping is represented by a ControlNet that produces animal images across mammals, amphibians, reptiles and birds following a 2D input pose condition ϕ_j^{proj} learned by minimizing the following objective:

$$\mathcal{L}_{ControlNet} = \mathbb{E}_{z_0, t, y_j, \phi_j^{proj}, \epsilon \sim \mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(z_t; t, y_j, \phi_j^{proj})\|^2 \right], \tag{1}$$

where $z_0 = x_j$. The above objective aims to learn a network ϵ_θ to estimate the noise added to an input image z_0 (or x_j) to form a noisy image z_t given time-steps t, text y_j and pose condition ϕ_j^{proj}. The network ϵ_θ is represented by the standard U-Net architecture of diffusion models (Stable Diffusion in this case) with a trainable copy of the U-Net's encoder attached to it using trainable zero convolution layers. We provide training details in Sec. F.

The trained TetraPose ControlNet can be used to generate pose-controlled images of tetrapods, including mammals, reptiles, birds, and amphibians. The model performs well for out-of-domain 2D pose inputs of animals not seen during training. It also performs well with inputs consisting of modified 2D poses that include extra appendages such as multiple heads, limbs, wings, and/or tails. While T2I diffusion models inherently provide a huge diversity to the generated outputs, the control module provides strong controlling signals to generate appropriate body parts in the right positions, alleviating the problem of T2I diffusion models producing inconsistent multi-view images when prompted using directional texts only.

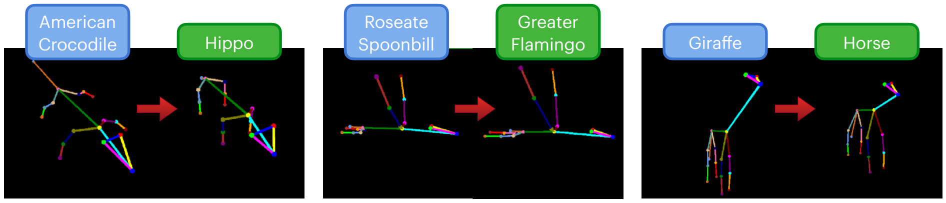

Figure 3: **Qualitative examples of pose editing using multi-agent LLM setup.** For each example, the green box denotes the desired animal, while the blue box is the animal retrieved from the 3D pose library by Finder LLM (π_F). We show the pose modification performed by the joint effort of Observer (π_O) and Modifier (π_M) for three instances.

3.2 3D Pose Generation aided by Multi-agent LLM

Generating a 3D pose based on a text is not trivial as text-to-pose is a many-to-many mapping. Existing 3D animal pose datasets are not diverse or vast enough to learn this mapping for a variety of animals. Thus we leverage LLMs which are pre-trained on expansive textual datasets, and thus can reason about anatomical proportions of various animals. We find that LLMs do not produce good 3D poses using only a text input, instead we use LLMs to adapt a input 3D pose to represent a novel animal. We created a limited library consisting of 16 animal 3D poses for this purpose.

Given a library of animals $\mathcal{B} = \{(y_i, \phi_i)\}_{i=1}^{n}$ consisting of 3D keypoint positions $\phi_i \in \Phi$ and animal names $y_i \in \mathcal{Y}$, we utilize a multi-agent LLM setup for creating a 3D pose for any desired animal y and pose description p. The agents include 1) Finder (π_F), 2) Observer (π_O), and 3) Modifier (π_M). Let the keypoint names representing any animal be $\mathcal{K}$ and let the bone sequence which defines the skeleton be $\mathcal{S}$. Given $\mathcal{K}$, the Finder selects the animal in $\mathcal{B}$ that is "anatomically closest" to the desired animal y as $(y_c, \phi_c) = \pi_F(y, \mathcal{B}, \mathcal{K})$. "Anatomically closest" is defined as the animal whose 3D pose will require minimal modifications/updates to represent y. Given the keypoint definitions $\mathcal{K}$, bone sequence $\mathcal{S}$, the desired animal name y, the animal y_c selected by π_F, and the pose description p, the Observer generates $\mathcal{O} = \pi_O(y_c, y, p, \mathcal{S}, \mathcal{K})$. $\mathcal{O}$ represents a plan describing which keypoints of y_c should be adjusted along with a set of instructions for the Modifier to implement the suggested adjustments to represent the 3D pose of the desired animal y in the described pose p. Based on the observations $\mathcal{O}$, the Modifier updates the 3D positions of the keypoints, ϕ_c, of the closest animal to $\phi = \pi_M(\phi_c, \mathcal{O})$. Thus we obtain the 3D keypoint positions ϕ of the desired animal y in the described pose p. We find that this multi-agent procedure is more stable and accurate than using a single LLM for pose generation (see Sec. B). We are able to represent diverse animals observed in nature using this setup. Fig. 3 presents examples of pose editing using our described setup. As ground truth text-to-3D poses for animals do not exist and the described task is a many-to-many problem, quantitative evaluation is difficult to obtain. Thus we conducted a user study to evaluate the efficacy of our method (details in Sec. 4). We describe the contents of our library $\mathcal{B}$ and the prompts to LLMs in detail in Sec. G and Sec. H.

3.3 Pose Editor and Shape Initializer

To facilitate easy creation and editing of 3D poses, we present a user-friendly tool to modify, add, or delete joints and bones. We also provide a method in this tool to automatically generate an initial shape based on the 3D skeleton using simple 3D geometries such as cylinders, cones, and ellipses. We use this shape to pre-train our NeRF before fine-tuning using diffusion based guidance. Details of this tool are presented in Sec. F.

3.4 Bringing Bones to Life

We want to create 3D animals given an input text y and 3D pose ϕ. We adopt the Score Distillation Sampling (SDS) method proposed in DreamFusion Poole et al. (2022), adapted for our TetraPose ControlNet. The SDS loss gradient can now be represented as:

$$\nabla_\eta \mathcal{L}_{\text{SDS}}(\theta, z = \mathscr{E}(g(\eta, c))) = \mathbb{E}_{t,c,\epsilon}\left[w(t)\left(\epsilon_\theta(z_t; t, y(c), \phi^{proj}(c)) - \epsilon\right)\frac{\partial z}{\partial \eta}\right], \quad (2)$$

Figure 4: **Comparison on generating animals observed in nature.** We compare with baselines which use T2I diffusion (with official open-source code) for the automatic generation of text-to-3D animals. Unlike the baselines, our method produces high quality anatomically consistent animals.

where η represents the trainable parameters of the NeRF, θ the frozen diffusion model parameters, c represents the sampled camera parameters, t is the number of time-steps, and $w(t)$ is a timestep-dependent weighting function. z denotes the latent encoded using encoder $\mathscr{E}$ for the image rendered from the NeRF $g(\eta, c)$ for camera c. $y(c)$ represents the directional text created based on camera c, while $\phi^{proj}(c)$ is the 2D projection of the 3D pose ϕ for camera c. Additionally, we also utilize an image domain loss weighted by the hyper-parameter λ_{RGB} which reduces flickering and produces more solid geometry (Fig.14):

$$\mathcal{L}_{RGB} = \lambda_{RGB} \cdot \mathbb{E}_{t,c,\epsilon} \left[w(t) \cdot \|g(\eta, c) - \mathscr{D}(\hat{z})\|^2 \right], \tag{3}$$

where $g(\eta, c)$ is the image rendered from the NeRF and $\mathscr{D}(\hat{z})$ is the denoised image decoded using decoder $\mathscr{D}$ from the denoised latent $\hat{z}$.

Since our TetraPose ControlNet is trained on much smaller number of images compared to Stable Diffusion, it loses diversity. To improve diversity and generation capability, we propose to use control scheduling and guidance scheduling. We observe that higher control scale provides strong signal for geometry modeling whereas higher guidance scale provides strong signal for appearance modeling. Since geometry is perfected in the initial stages and appearance in the latter, we propose reduction of control scale and increase of guidance scale over training iterations. This helps us create out-of-domain assets with significant style variety (see Fig. 8). Our strategy is formulated as:

$$\text{control\_scale} = \cos\left(\frac{\pi}{2} \cdot \frac{\text{train\_step}}{\text{max\_step}}\right) \cdot (\text{control}_{max} - \text{control}_{min}) + \text{control}_{min}, \tag{4}$$

$$\text{guidance\_scale} = \frac{\text{train\_step}}{\text{max\_step}} \cdot (\text{guidance}_{max} - \text{guidance}_{min}) + \text{guidance}_{min}, \tag{5}$$

where train_step is the current training step and max_step is total training iterations. The variables control_{max}, control_{min}, guidance_{max}, guidance_{min} are hyperparameters. We show that a linear scheme is better for guidance scheduling, while cosine is better for control scheduling in Sec. B.

4 Experiments

In this section, we compare YOUDREAM against various baselines and evaluate the effect of various components of our method. We show qualitative comparison with text-to-3D methods which are guided by T2I diffusion models for common animals observed in nature. We also compare with MVDream – which uses a (text + camera)-to-image diffusion model trained on 3D objects. It should be noted that our method does not use any 3D objects for training, yet is able to deliver geometrically consistent results. We conduct a user study to quantitatively evaluate our method against these baselines. We also compute CLIP score following previous work Shi et al. (2023), shown in Sec. I. Additionally, we present ablations over the various modules that constitute YOUDREAM.

Generating animals observed in nature. In Fig. 4, we compare our method against 3DFuse Seo et al. (2023), Fantasia3D Chen et al. (2023a), and HiFA Zhu et al. (2023) for generating common animals. HiFA and Fantasia suffer from anatomical inconsistency, while 3DFuse is more consistent in some cases due to the use of depth control. However, 3DFuse is highly dependent on the point cloud prediction leading to the generation of implausible geometry, for example in the case of elephant and T-Rex. It should be noted that generating results using Fantasia3D required extensive parameter tuning, which has also been indicated by the authors in their repository. All results are generated using the same seed 0 for fair comparison. We use the default hyperparameter settings of each baseline except Fantasia3D. The text ", full body" is appended at the end of the prompt for all baselines, as we observed that the methods generate truncated animals in many cases. We generate common animals using our fully automated pipeline, where we use LLM for pose editing sourced by a library of 3D poses. Tiger is generated by our multi-agent LLM based on a German Shepherd, northern cardinal is made from an eagle, while elephant and Tyrannosaurus rex are part of our library. In all cases, our method visibly outperforms baselines in terms of perceptual quality and 3D consistency.

Generating unreal creatures. A major advantage of our pipeline is that it can be easily used to generate non-existent creatures, especially those not explainable through text. These can be generated robustly using our method when the user provides a skeleton of their concept. We use our pose editor tool to generate the results shown in Fig. 1, where YOUDREAM produces stunning unreal creatures. We show the pose controls we use in Sec. F. Notably, MVDream Shi et al. (2023) struggled to follow the textual prompt as such creatures are not represented in existing 3D datasets; producing incorrect results such as "Wampus cat", a cat-like creature in American folklore, with three legs instead of six in Fig. 1 row 5. In some aspects HiFA attempted to follow the prompt (owing to its usage of T2I Stable Diffusion) such as producing a couple of tentacles in Fig. 1 row 1 and more than one head in row 3, but produces geometrically inconsistent results in all cases. Again we use seed 0, default hyperparameter settings for baselines and append ", full body" at the end of the prompts except for 'golden ball'.

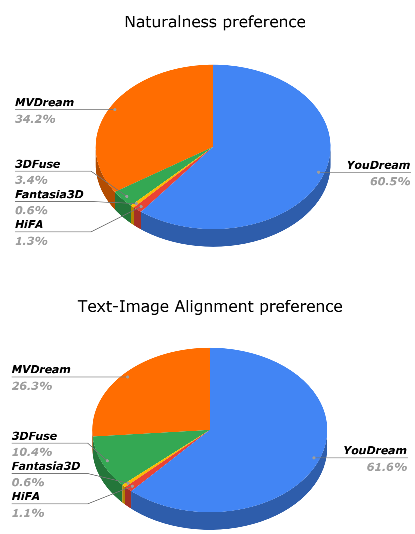

Figure 5: **User Study.** User preferences on 1) Naturalness and 2) Text-Image alignment averaged over 32 participants and 22 text-to-3D generated assets reveals the superiority of our proposed method.

Subjective Quality Analysis. We conducted a voluntary user study[2] with 32 participants to subjectively evaluate the quality of our 3D generated assets. The participants were shown side-by-side videos of assets generated using the same prompt input by YOUDREAM(ours), HiFA, Fantasia3D, 3DFuse, and MVDream, and were asked to select the best model under the categories - 1) Naturalness and 2) Text-Image alignment. The participants were instructed to judge naturalness on the basis of geometrical and anatomical consistency/correctness, perceptual quality, artifacts, and details present in the videos.

[2] This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB), University of Texas, Austin, under FWA No. 00002030 and Protocol No. 2007-11-0066.

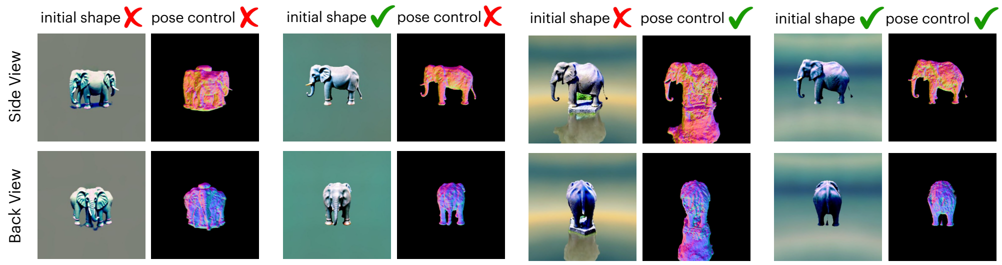

Figure 6: **Ablation over the effect of initial shape and pose control.** The initial shape helps in producing clean geometry, while the pose control helps to maintain 3D consistency.

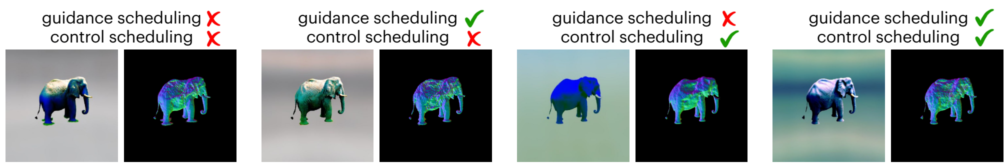

Figure 7: **Ablation over scheduling techniques.** Using either guidance or control scaling produces unnatural color, using neither produces artifacts such as grass at feet owing to lower diversity of ControlNet compared to Stable Diffusion.

Text-image alignment preference was self-explanatory. A total of 22 prompts and their corresponding 3D assets generated using each model were shown to each participant, thereby accumulating a total of 1408 user preferences. Of the 22 prompts, 13 involved naturally existing animals while the remaining 9 included unreal and non-existent animals. The collected user preferences are shown in Fig.5. We observe a 60-62% user preference in both the preference categories for our model, strongly indicating the superior robustness and quality of YOUDREAM.

We also tested the efficacy of our multi-agent LLM based pose generator via a subjective study. We request 16 novel 3D poses of different animals from the multi-agent LLM which uses 16 pre-defined animal poses in our animal pose library. The requested animals were such that there was a high chance of using each reference animal pose in the library. The participants were shown paired videos of rotating 3D poses, consisting of the pose taken from the library (left-side video, 'reference animal') and the novel pose (right-side video, 'requested animal') generated. Since the participants were not experts in animal anatomy, they were also provided multi-view images of each animal under their video. They were asked to mark 'Yes' or 'No' for the question: "If this 3D pose <reference pose video> represents 'reference animal' in 'reference pose' pose. Could this 3D pose <generated pose video> represent 'requested animal' in 'requested pose' pose?" The study consisted of the same 32 participants and each subject voted for all the 16 novel poses. Subjects agreed that the generated pose correctly represents 'requested animal' 91% of the time. Kangaroo (standing pose) generated by the multi-agent LLM using the pose of T-Rex (standing pose) received the lowest agreement among all pairs with 8 out of 32 votes being 'No'. Detailed description of the pose library, the generated poses, and particulars of the human study are provided in Sec. J.

Ablation. We present ablation over the effect of using initial shape and pose control in Fig. 6. Without pose control refers to using vanilla Stable Diffusion. Without using initial shape or control, the Janus head problem occurs. With initial shape but without pose control, the geometry improves but still sees the appearance of another head on the elephant's backside. Using pose control without initial shape produces visibly good results, however using both initial shape and pose control results in much cleaner geometry.

In Fig. 7 we show the effect of our scheduling strategies. Without guidance or control scaling, the result has grassy texture at the feet which could be owing to seeing most elephants on grass during TetraPose ControlNet training on limited animal pose data. Using only one kind of scheduling produces incorrect color, showing that both scaling techniques go hand-in-hand.

5 Conclusion

We presented YouDream, a method to create anatomically controllable and geometrically consistent 3D animals from a text prompt and a 3D pose input. Our method facilitates the generation of diverse creative assets through skeleton control, which cannot be expressed through language and is difficult to provide as guidance image, especially for unseen creatures. Additionally, we presented a pipeline for automatic generation of 3D pose for animals commonly observed in nature by utilizing a multi-agent LLM setup. Our 3D generation process enjoys multi-view consistency by utilizing a 3D pose as a prior. We quantitatively outperform prior work in terms of "Naturalness" and "Text-Image Alignment" as evidenced in the user study.

6 Acknowledgment

The authors would like to thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing compute resources that have contributed to the research results reported in this paper. URL: http://www.tacc.utexas.edu.

References

Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023.

Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.

Prianka Banik, Lin Li, and Xishuang Dong. A novel dataset for keypoint detection of quadruped animals from images. *arXiv preprint arXiv:2108.13958*, 2021.

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023.

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.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22246–22256, 2023a.

Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1148–1156, 2023b.

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.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scene as blender code. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

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.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022.

Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3d: Learning articulated 3d animals by distilling 2d diffusion. *arXiv preprint arXiv:2304.10535*, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. *arXiv preprint arXiv:2401.02400*, 2024.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

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

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference papers*, pages 1–8, 2022.

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.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

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

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.

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.

Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17542–17552, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Hyeonsu Kim, Jaehoon Ko, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv preprint arXiv:2311.15475*, 2023.

Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3d-gpt: Procedural 3d modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023.

Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, 131(10):2623–2634, 2023a.

Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023b.

Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4853–4862, 2023.

Chun-Han Yao, Amit Raj, Wei-Chih Hung, Michael Rubinstein, Yuanzhen Li, Ming-Hsuan Yang, and Varun Jampani. Artic3d: Learning robust articulated 3d shapes from noisy web image collections. *Advances in Neural Information Processing Systems*, 36, 2024.

Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023.

Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Jiayuan Fan, Gang Yu, Taihao Li, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *arXiv preprint arXiv:2311.17618*, 2023.

Zeqing Yuan, Haoxuan Lan, Qiang Zou, and Junbo Zhao. 3d-premise: Can large language models generate 3d shapes with sharp features and parametric control? *arXiv preprint arXiv:2401.06437*, 2024.

Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Daniel Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7124–7132, 2024a.

Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. In *European Conference on Computer Vision*, pages 668–685. Springer, 2022.

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.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024b.

Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *The Twelfth International Conference on Learning Representations*, 2023.

Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018.

Appendix

A More Results

We show our method's performance for varied styles in Fig. 8. Even though our ControlNet is trained on images of animals in the wild, YOUDREAM produces assets with significant style alteration. This is attainable because of our control scheduling and guidance scheduling approach, which ensures consistent geometry to be formed in initial iterations with higher control scale and style being finalized during the later iteration with higher guidance scale.

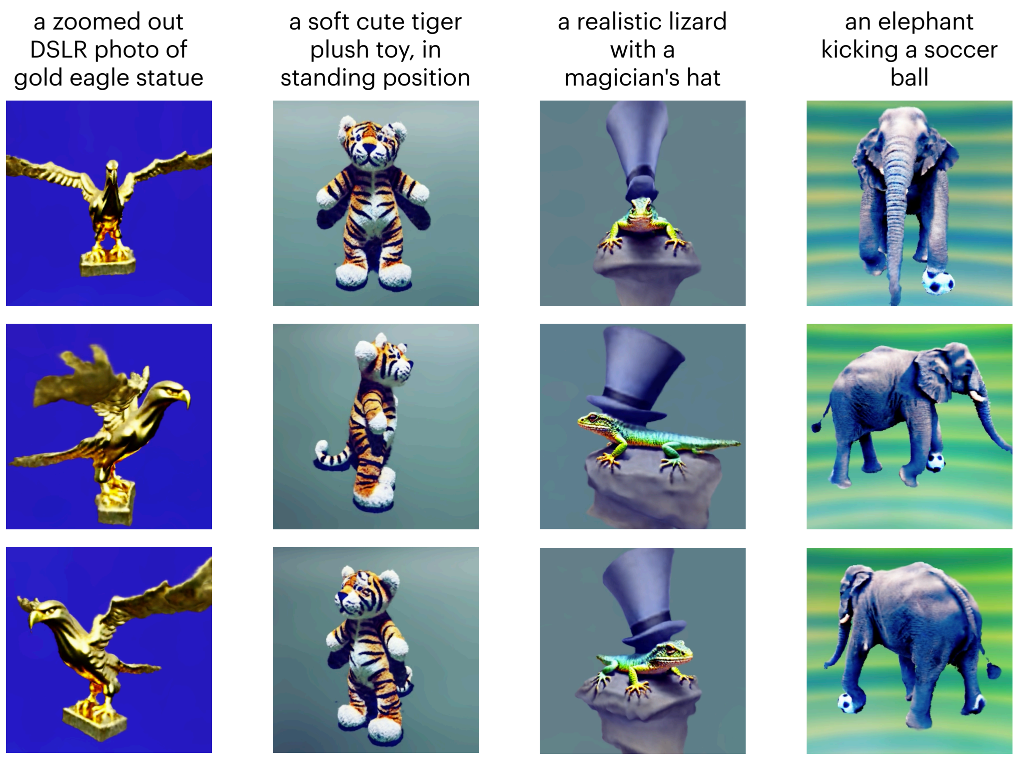

Figure 8: **Results on compositional and style prompts.** We show our method performs well while generating animals with style alterations or object interactions.

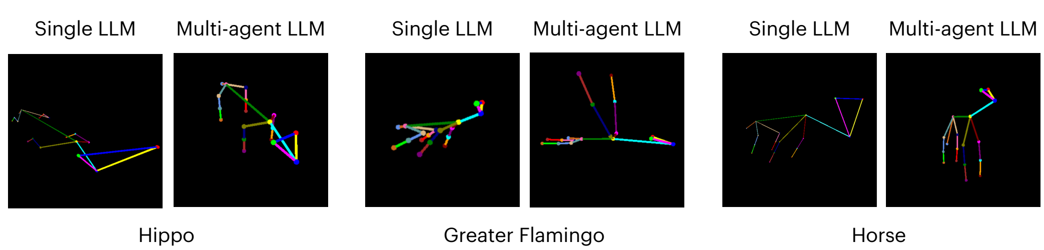

Figure 9: **Pose generation using single LLM vs our multi-agent LLM setup.** For "Hippo", "Greater Flamingo", and "Horse", we show a 2D view of the 3D pose generated by a single LLM compared to our multi-agent setup.

B Additional Ablations

In Fig. 9 we show that using a single LLM agent performs much worse in generating 3D poses compared to our multi-agent setup which includes Finder, Observer and Modifier GPTs.

Figure 10: **Comparison with OpenPose ControlNet for generating animals.** OpenPose ControlNet produces the animal in the prompt for "Horse" and "Baboon", but either does not follow control or makes unnatural anatomy. For "Gazelle" a meaningless image is produced.

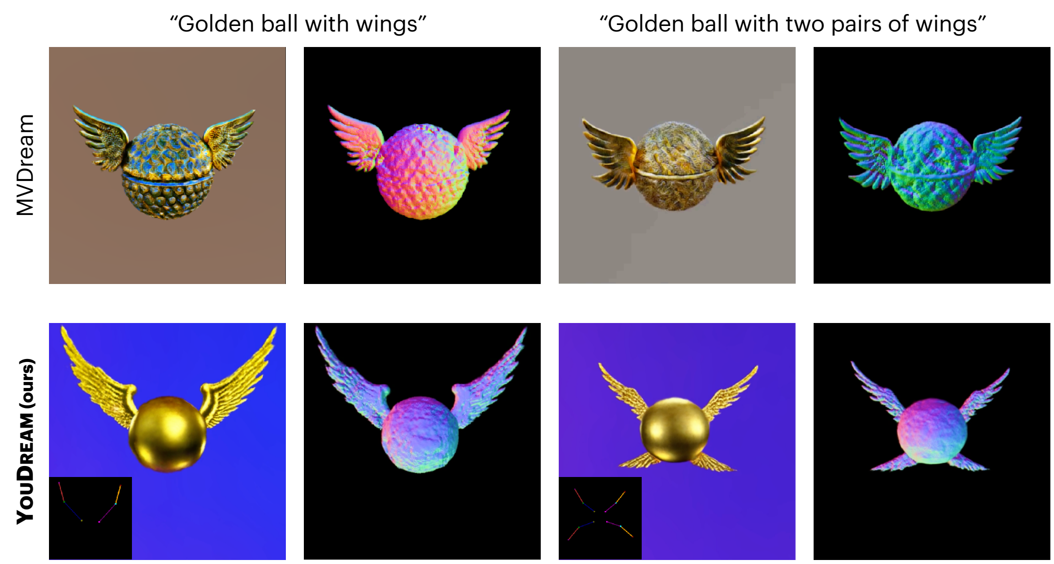

Figure 11: **Toy example showing inefficacy of text prompt.** We show that pose control helps to add the additional wings at the desired location. MVDream makes two wings for "Golden ball with two pairs of wings" with differently shaped wings compared to "Golden ball with wings".

Since our ControlNet guided 3D generation pipeline can produce out-of-domain animals well, the question arises if OpenPose ControlNet can be utilized to generate animals. We show in Fig. 10 that OpenPose ControlNet produces artificial looking images for animals. We use a "human on all fours" image to obtain the pose for OpenPose and generate a similar keypoint orientation for our TetraPose format. Even though the pose is unnatural for animals, with hips and shoulder very close to spine, TetraPose ControlNet produces clean images following the pose.

A toy example based on golden ball with wings is presented in Fig. 11 to show that text by itself can be ambiguous to convey meaning. When prompted for two wings, MVDream produces a modified pair of wings, whereas YouDream follows the user pose control to produce four wings. **YouDream** performs significantly better for many prompts involving real animals such as pangolin and giraffe.

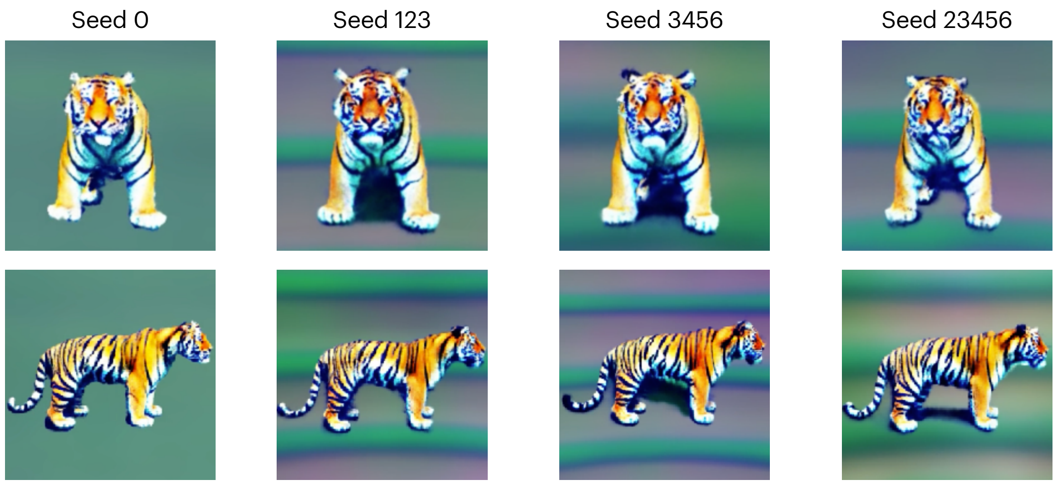

Figure 12: **Variation with seed.** Our method is robust across seeds and generates slightly different faces and stripes for various seeds.

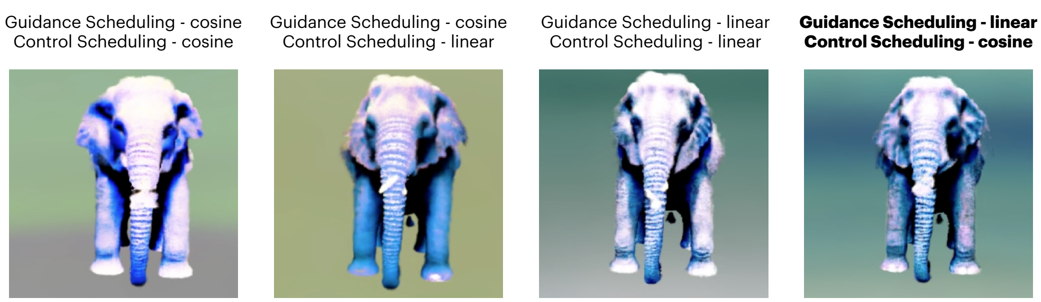

Figure 13: **Comparison of various scheduling techniques.** Using cosine strategy for both produces oversaturation, while using cosine strategy for guidance scheduling and linear for control scheduling produces oversmooth textures at the legs. Results of using both linear scheduling is closest to our strategy, but is lesser textured (notice feet and ears).

Fig. 15 shows results for both the animals generated using MVDream and YouDream. Even though MVDream is a 3D aware model, it still produces artificial looking results in many cases. While results generated using YouDream are much more natural perceptually and contain realistic textures found in the respective animals.

We show that our method does not require seed tuning for generating consistent results in Fig. 12. Variation in textures and shapes can be seen across seed.

In Fig. 13 we show the effect of different guidance and control scheduling strategies. Note that for all, guidance scale increases while control scale reduces.

We show that not using $\mathcal{L}_{RGB}$ loss produces holes and flickering in generated assets. We show the normals for elephant and tiger for this purpose.

Comparison with 3D Animal Model. We compared our method against 3DFauna Li et al. (2024), a 3D animal reconstruction method based on image inputs. Given an input image 3DFauna failed to capture high-frequency details and follow the input image (see tail and snout in Fig. 16), whereas our method produced a highly detailed animal given the input pose and text, which closely followed the input pose control.

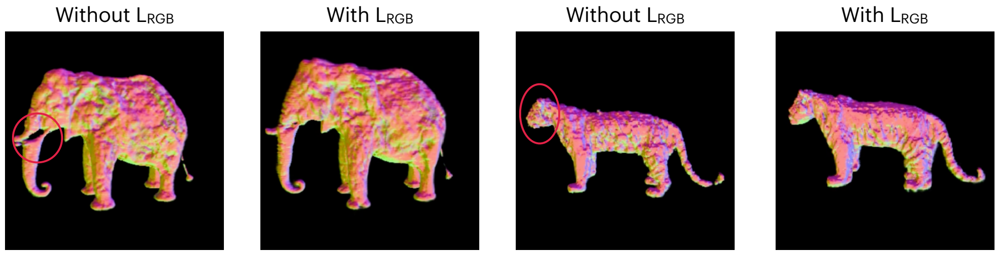

Figure 14: **Effect of using $\mathcal{L}_{RGB}$.** Not using $\mathcal{L}_{RGB}$ results is hollow geometry and flickering. The chin of the tiger appears and disappears based on view, a view where the chin has disappeared has been chosen.

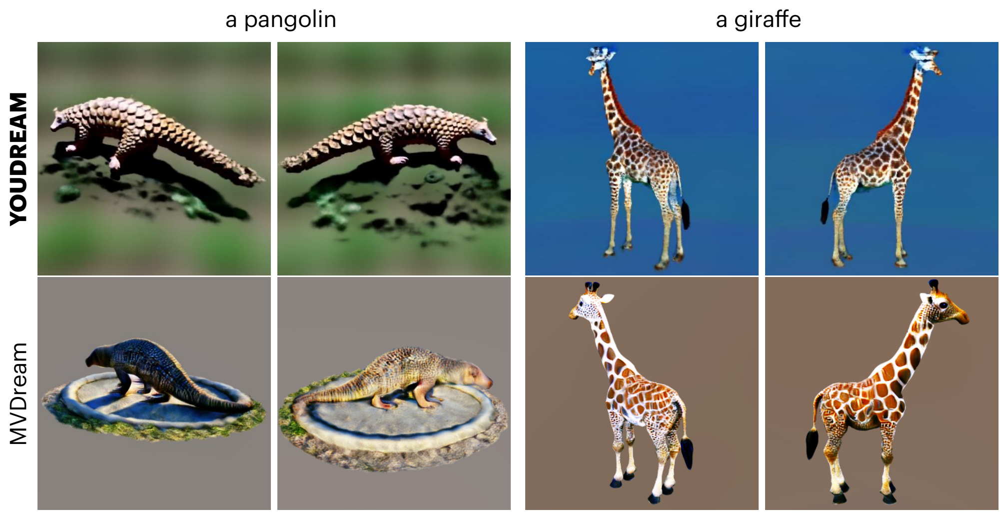

Figure 15: **More Comparison with MVDream.** We compare our method with MVDream for simple prompts. MVDream results are clearly missing the texture of the scaly body of the pangolin, while their giraffe has a toy-like geometry and hence unnatural. In contrast YOUDREAM produces very realistic results.

C Comparison with more text-to-3D baselines

We also compare with other text-to-3D generative methods guided by T2I models. These include Stable Dreamfusion Tang (2022), ProlificDreamer Wang et al. (2024), and LucidDreamer Liang et al. (2023). In Fig. 17, we show that all these methods suffer geometric and anatomic inconsistencies, as well as fail to capture the text faithfully.

D Exploring severely Out-of-Domain cases

We explore generating animals well out-of-domain with respect to our animal library (see Sec. G). We show in Fig. 18 that we can generate “clownfish" and “four-legged tarantula" without any human intervention using our fully automatic pipeline comprising of the multi-agent LLM pose editor and the 3D generation pipeline. Our multi-agent LLM setup has been explored in the context of four-limbed animals and generating more appendages is a direction of future work.

E Scaling to higher dimension NeRF

In Fig. 19 , we show that we can scale 3D generation to a higher dimension NeRF without any changes in hyperparameters. It can be observed that scaling to the larger NeRF improves the sharpness of the asset considered and results in crisper textures.

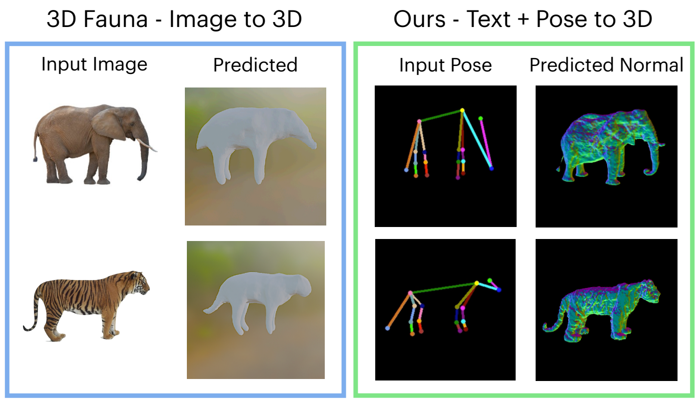

Figure 16: **Comparison with 3DFauna.** Our method produces more detailed geometry compared to the baseline.

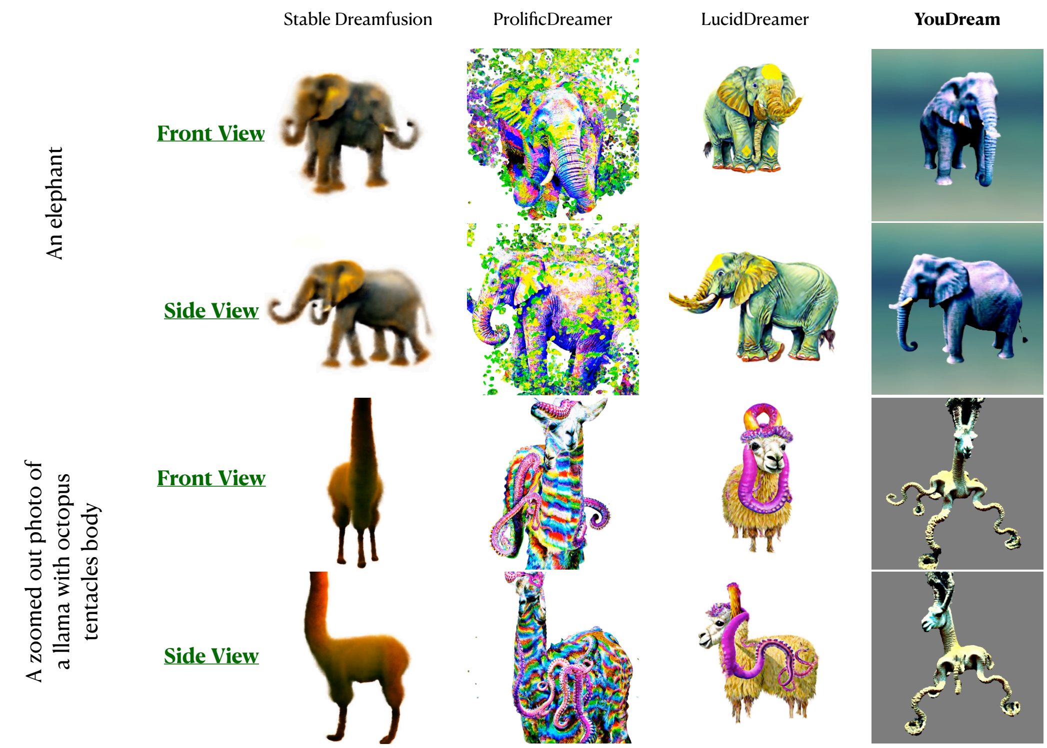

Figure 17: **Comparison with additional prior art methods.** Even though LucidDreamer performs better than Stable Dreamfusion and ProlificDreamer, it shows the same failures as discussed in the main paper.

F Implementation details

Poses used to generate 3D animals in main paper. Fig. 20 shows the 2D views of 3D poses used to generate the 3D animals in main paper.

TetraPose ControlNet Training: We used annotated poses from the AwA-pose Banik et al. (2021) and Animal Kingdom Ng et al. (2022) datasets to train ControlNet in a similar way as the original paper, which uses stable diffusion version 1.5. AwA-pose consists of 10k annotated images covering 35 quadruped animal classes, while Animal Kingdom provides 33k annotated images spanning 850 species, including mammals, reptiles, birds, amphibians, fishes and insects. From a combined set of 43k samples, we carefully selected a subset including only mammals, reptiles, birds, and amphibians. We also eliminated any sample having less than 30% of its keypoints annotated. The curated dataset consists of 13k annotated samples. To increase diversity in learning, and to improve

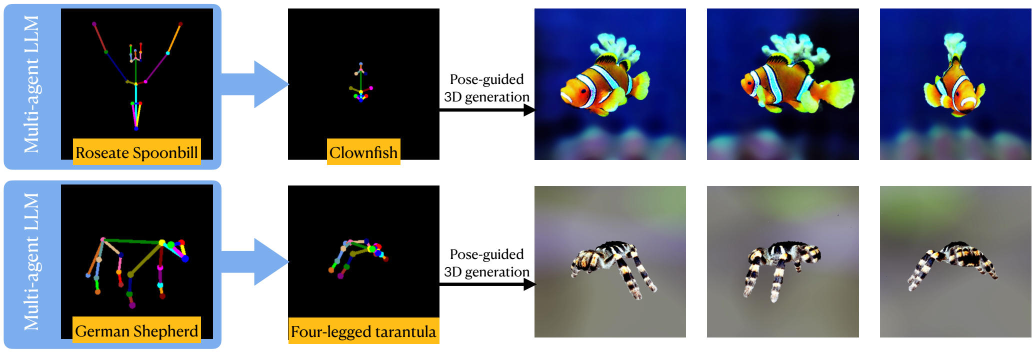

Figure 18: **Generating more OOD assets through automatic pipeline.** Using our multi-agent LLM setup we first generate the 3D poses of “clownfish" and “four-legged tarantula". We then use the produced 3D poses to guide our 3D generation. We observe that the multi-agent LLM pose editor chooses “Roseate Spoonbill" as the base 3D pose to be modified into “Clownfish", and “German Shepherd" is chosen for modifying to “four-legged tarantula".

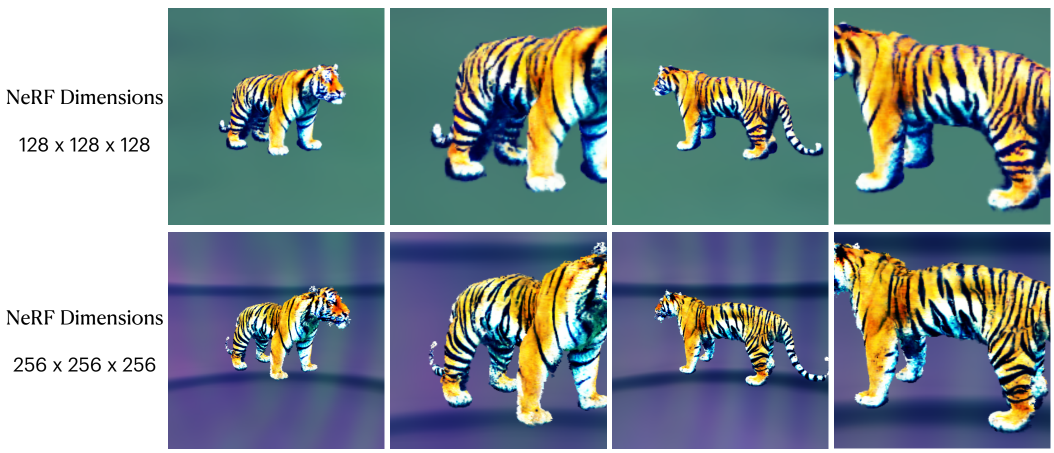

Figure 19: **Increasing NeRF dimensions.** On increasing each NeRF dimension by 2$\times$ we generate a sharper and cleaner 3D asset for the prompt “a tiger" without any change in hyperparameters.

test-time generation at any scale and transformations, we used a combination of data augmentation strategies consisting of random rotations, translations, and scaling while training so as to handle highly varied and heavily occluded 2D pose samples during 3D generation. The model was trained over 229k iterations with a batch size of 12, a constant learning rate of $1e^{-5}$, on a single Nvidia RTX 6000. The model converged after around 120k iterations and would not overfit even up to 200k iterations, owing in part to the augmentation strategy.

3D Pose editing and Shape generation: We used the following 18 keypoints to represent every quadruped: left eye, right eye, nose, neck end, 4 $\times$ thighs, 4 $\times$ knees, 4 $\times$ paws, back end, and tail end. For the upper limbs of birds, i.e. wings, their front - thighs, knees, and paws are defined in accordance with how their upper limbs move. The user can begin with any initial pose from the animal library and modify its keypoints using the Balloon Animal Creator Tool. This tool was developed using THREE.js and can be run on any web browser. The tool provides buttons for the following functions: 1) add extra head. 2) add extra limb - front, 3) add extra limb - back, and 4) add extra tail. After appropriate modification of the pose the user can press the button to *create mesh* around bones. This button press invokes calls to various functions defined to create each body part, based on their natural appearances using simple mesh components such as ellipses (eyes and torso), cylinders (neck, tail, and limbs), and cones (nose). The combined mesh and the corresponding keypoints can be downloaded by clicking the *Export Mesh* and *Save Keypoints* button. An example of this process used for creating the three headed dragon using the Balloon Animal Creator tool is depicted in Fig. 21.

Figure 20: **Snapshots of 3D poses used for generating objects in the main paper.** For a 2D view of each object, we show the corresponding 2D view of the 3D pose.

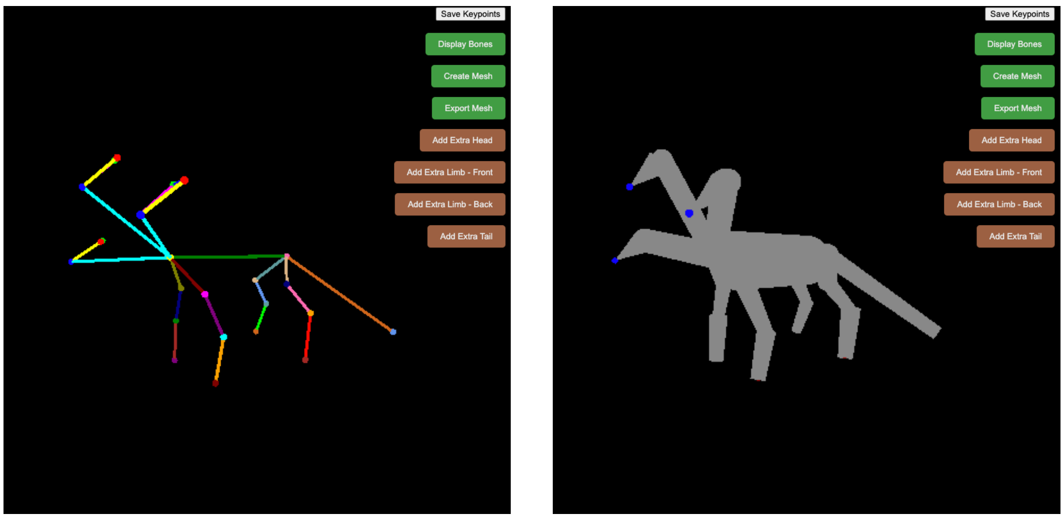

Figure 21: **3D Pose editing and Shape generation.** We show snapshots of our 3D pose creator tool with all functionalities.

Mesh depth guided NeRF initialization: The mesh downloaded in the previous step was used to provide depth maps to the pre-trained depth guided ControlNet, which produces the gradient loss by SDS, which in turn is used to pre-train the NeRF. The pre-training helps achieve a reasonable initial state for the NeRF weights, which can then be refined in the final pose-guided training stage. The diffusion model was pre-trained for 10,000 iterations using the Adam optimizer with a learning rate of $1e-3$ and a batch size of 1. During training, the camera positions were randomly sampled in spherical coordinates, where the radius, azimuth, and polar angle of camera position were sampled from [1.0, 2.0], [0, 360], and [60, 120].

Pose-guided SDS for NeRF fine-tuning: Finally, we fine-tune the NeRF using the pre-trained ControlNet to provide 2D pose guidance to SDS. The gradients computed using the noise residual from SDS were weighted in a similar manner as DreamFusion, where $w(t) = \sigma_t^2$ and t was annealed using $t = t_{max} - (t_{max} - t_{min})\sqrt{\frac{iter}{total\_iters}}$. We set t_{max} to be 0.98, t_{min} to be 0.4. Similar to the previous stage, we trained the model over $total\_iters = 10,000$ using the same settings for the optimizer. Using cosine annealing, we reduced the $control_{scale}$ from an initial value of 1 to a final value of 0.2, while updating $guidance_{scale}$ linearly from $guidance_{min} = 50$ to $guidance_{max} = 100$. These settings helped reduce the impact of ControlNet gradually over the training process, while improving quality by gradually increasing strength of classifier-free guidance. The camera positions were randomly sampled as in stage 1, as were the radius, azimuth, and polar angle of the camera. λ_{RGB} was set to 0.01. The 3D avatar representation renders images directly in the RGB space of $\mathbb{R}^{128\times128\times3}$. We use Instant-NGP Müller et al. (2022) as the NeRF representation. The pre-training stage, if used, takes less than 12 minutes to complete, while the fine-tuning stage takes less than 40 minutes to complete on a single A100 40GB GPU.

Computational Resources: All the experiments pertaining to YOUDREAM and 3DFuse were run on Nvidia A100 40GB GPU. Few experiments for MVDream and all experiments of HiFA required running on A100 80GB GPU, while all experiments for Fantasia3D were run on 3xA100 40GB GPUs.

G Animal Library

Our animal library $\mathcal{B}$ contains a total of 16 animal/pose combinations:

- Giraffe
- Elephant
- German Shepherd
- Eagle - sitting
- Eagle - flying
- American Crocodile
- Tree Frog
- Roseate Spoonbill - sitting
- Roseate Spoonbill - flying
- Raccoon - standing on four legs
- Raccoon - standing on two legs
- T-Rex
- Lizard
- Tortoise
- Bat
- Otter

All common animals results shown in this paper are either using these 3D poses or poses modified from one of these by our multi-agent LLM. The library entries are chosen intuitively such that each has significant anatomical variation from the others so as to cover the large range of variety observed in the animal kingdom.

H Multi-agent LLM Implementation Details and Scope

We use the recently released "GPT-4o" API of OpenAI with max_tokens as 4096 and temperature as 0.9. The keypoints are represented as a dictionary in JSON format and converted to a string to be appended to the text prompts of the observer and modifier LLMs. The observer LLM is instructed using the system prompt about details of the 3D coordinate space and relations among the various keypoints. The bone sequence which represents the connections between the various keypoints is also provided as a list to the observer GPT's prompt to reason about relative anatomy based on bone lengths. Finally, the multi-agent LLM outputs a keypoint dictionary in the same format as provided to it. The Multi-agent LLM is able to generate various animals by taking reference from a set of 16 3D animal poses. However, this setup can generate poses for animals that are well out-of-domain of these 16 animals as shown in Fig. 18. The Multi-agent LLM supports generating animals that can be represented using four limbs. Generating more than four-limbed animals such as insects using the LLM setup is a direction of future work. We open-source this setup along with our project code.

I Evaluation using CLIP score

Based on the user study, it is clear that users majorly prefer either MVDream or YOUDREAM. Hence we also compute CLIP similarity score for each of the two methods as the average CLIP score over 9 views of each of the 22 prompts used for the user study. Table 1 shows that our method outperforms MVDream based on the CLIP similarity score. We use the ViT-B/32 model for evaluation.

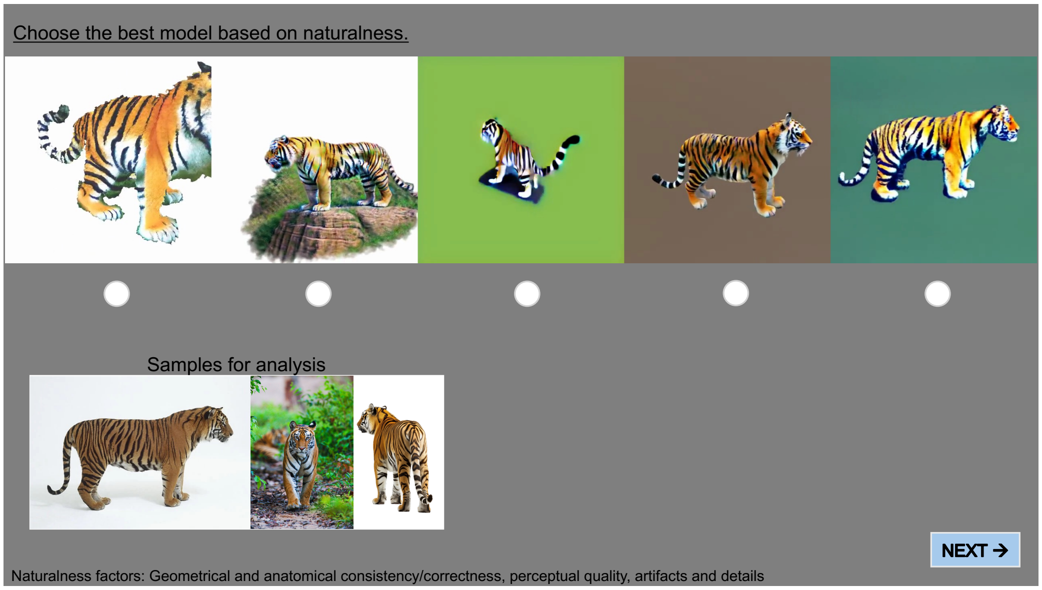

Figure 22: **User Study Interface** for *Naturalness preference*: A snapshot of the interface displaying rotating videos of the results generated by the five chosen models. The user was provided with sample images of the real animal for analyzing anatomical consistency.

	MVDREAM	YOUDREAM
CLIP Score ↑	29.78	**30.86**

Table 1: **CLIP similarity score** comparison for MVDream and YOUDREAM

J User Study Details

Graduate students at the University of Texas at Austin volunteered for participating in the user study. All the information regarding the preferences requested in the study, the judgement criteria, and operating the interface were provided at the beginning of the study. A consent form documenting the purpose of the study, risks involved in the study, duration of the study, compensation details, and contact details for grievances was signed by each user before the beginning of their study session. Screenshots of the user study interfaces are shown in Fig. 22 and Fig. 23. The 22 prompts used for generating the 3D assets used in the user study are as follows:

1. A giraffe
2. A lizard
3. A raccoon standing on two legs
4. A tiger
5. A lion
6. A red male northern cardinal flying with wings spread out
7. A roseate spoonbill flying with wings spread out
8. A Tyrannosaurus rex
9. A pangolin
10. A bear walking
11. A horse
12. A mastiff
13. A soft cute tiger plush toy, in standing position

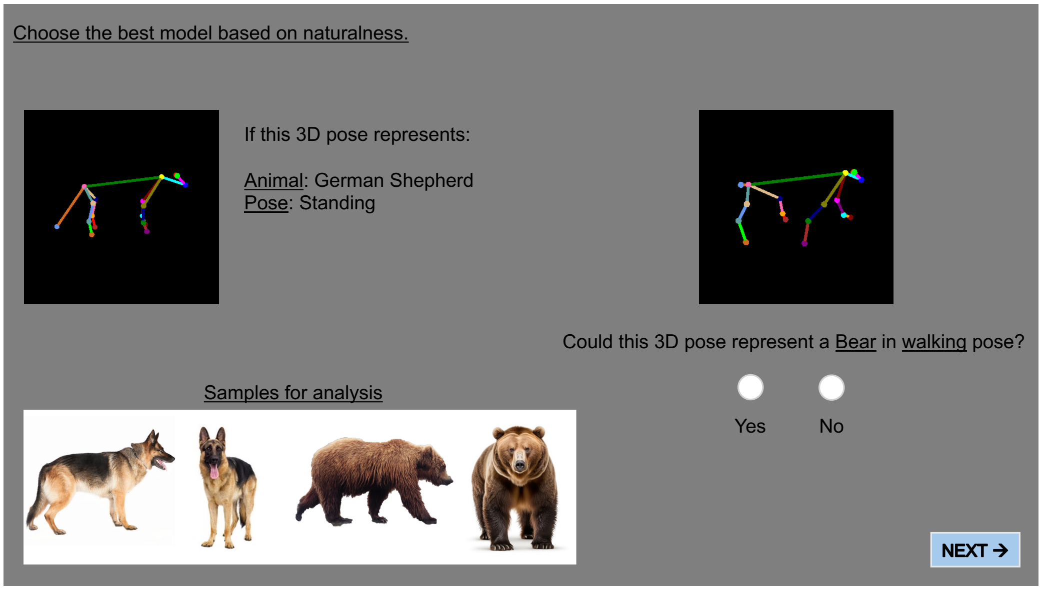

Figure 23: **User Study Interface** for *Generated Pose Preference*: A snapshot of the interface displaying rotating pose video of the 'reference animal' (left side of the interface) used by the multi-agent LLM for generating a 3D pose of 'requested animal' (right side of the interface) in 'requested pose'. The user was provided with real samples of the 'reference animal' and the 'requested animal' (one side-view and one front-view each for better anatomical analysis.

14. An elephant standing on concrete
15. A dragon with three heads separating from the neck
16. A realistic mythical bird with two pairs of wings and two long thin lion-like tails
17. Golden ball with wings
18. A six legged lioness, fierce beast, pouncing, ultra realistic, 4k
19. A giraffe with dragon wings
20. A zoomed out photo of a llama with octopus tentacles body
21. A zoomed out DSLR photo of a gold eagle statue
22. Golden ball with two pairs of wings

K Animation using Pose Sequence

YOUDREAM can also be used to generate animated videos by generating 3D assets for every pose from a pose sequence. In Fig. 24 we show frames chosen from a pose sequence and the corresponding render of their generated 3D mesh. Despite this, generating a longer animation sequence using YOUDREAM would be a highly resource expensive and time consuming task. We hope this work will inspire further exploration of efficient methods for controlled animation.

L Limitations and Discussion

While our method produces high quality anatomically consistent animals, the sharpness and textures can be improved by utilizing a number of tricks used by recent papers. We use a 128×128 NeRF, while our baseline HiFA uses 512×512, while MVDream uses 256×256. We use a smaller NeRF for the sake of lower time complexity compared to baselines. Other tactics such as using DMTet or regularization techniques are also plug-and-play for our method and may improve sharpness.

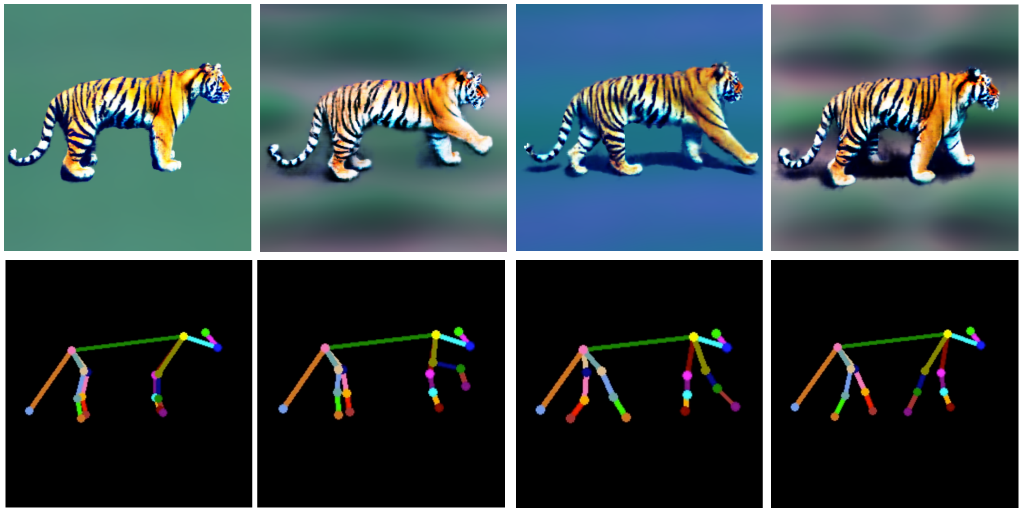

Figure 24: **Animation.** Bottom-row: Sampled pose frames from a pose sequence of a tiger walking. Top-row: Camera captured image of the 3D mesh corresponding to the view of the 3D pose shown below it in the bottom-row.

We show several diverse examples of automatically generating common animals found in nature. However there could exist unusually shaped animals whose 3D pose cannot be satisfactorily generated using our multi-agent LLM setup. In these cases, manual editing of 3D pose might be required over the LLM generated 3D pose. However, we believe our pose editor tool is highly interactive and user-friendly, thus requires very low human effort to modify poses.

Broader Impact. AI generated art has been widely used in recent times. YouDream enables artists to gain more control over their creations, thus making the process of content creation easier. As our method uses Stable Diffusion, it inherits the biases of that model. TetraPose ControlNet training uses existing open-source animal pose datasets instead of internet scraped images, hence avoiding any copyright issues.

Licenses

URL	Citation	License
`https://github.com/JunzheJosephZhu/HiFA`	Zhu et al. (2023)	Apache License 2.0
`https://github.com/KU-CVLAB/3DFuse`	Seo et al. (2023)	N/A
`https://github.com/Gorilla-Lab-SCUT/Fantasia3D`	Chen et al. (2023a)	Apache License 2.0
`https://github.com/bytedance/MVDream`	Shi et al. (2023)	MIT License
`https://github.com/ashawkey/stable-dreamfusion`	Tang (2022)	Apache License 2.0
`https://github.com/thu-ml/prolificdreamer`	Wang et al. (2024)	Apache License 2.0
`https://github.com/EnVision-Research/LucidDreamer`	Liang et al. (2023)	MIT License
`https://github.com/lllyasviel/ControlNet`	Zhang et al. (2023)	Apache License 2.0
`https://github.com/prinik/AwA-Pose`	Banik et al. (2021)	MIT License
`https://github.com/sutdcv/Animal-Kingdom`	Xu et al. (2023)	N/A

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:[NA]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The experiments conducted in this paper do not have either a ground truth or a well-defined metric for evaluation. We have reported CLIP score following prior work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.