MAS: Multi-view Ancestral Sampling for 3D Motion Generation Using 2D Diffusion

Roy Kapon, Guy Tevet, Daniel Cohen-Or and Amit H. Bermano
Tel Aviv University
roykapon@mail.tau.ac.il

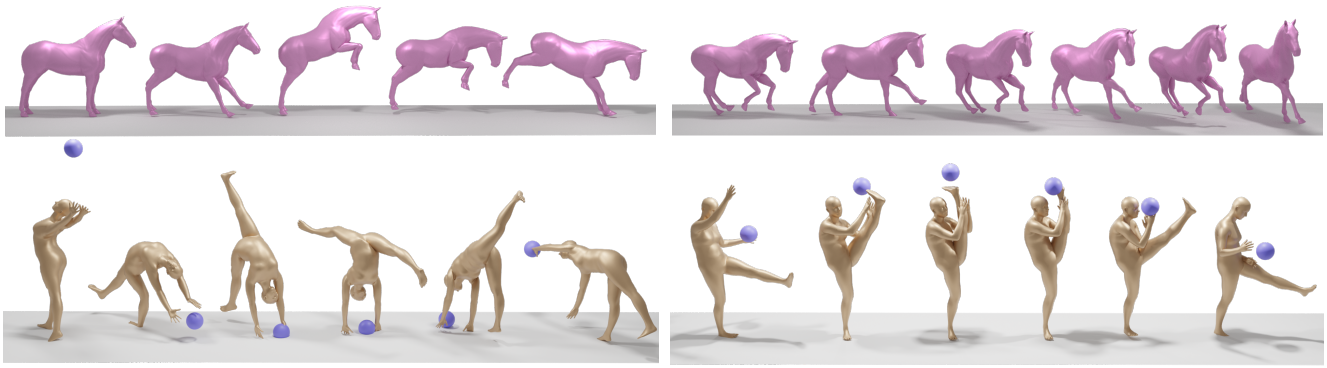

Figure 1. 3D motions generated by Multi-view Ancestral Sampling (MAS) — each one using a different initial noise. Our method generates novel 3D motions using a 2D diffusion model. As such, it enables learning intricate 3D motion synthesis solely from monocular video data.

Abstract

We introduce Multi-view Ancestral Sampling (MAS), a method for 3D motion generation, using 2D diffusion models that were trained on motions obtained from in-the-wild videos. As such, MAS opens opportunities to exciting and diverse fields of motion previously under-explored as 3D data is scarce and hard to collect. MAS works by simultaneously denoising multiple 2D motion sequences representing different views of the same 3D motion. It ensures consistency across all views at each diffusion step by combining the individual generations into a unified 3D sequence, and projecting it back to the original views. We demonstrate MAS on 2D pose data acquired from videos depicting professional basketball maneuvers, rhythmic gymnastic performances featuring a ball apparatus, and horse races. In each of these domains, 3D motion capture is arduous, and yet, MAS generates diverse and realistic 3D sequences. Unlike the Score Distillation approach, which optimizes each sample by repeatedly applying small fixes, our method uses a sampling process that was constructed for the diffusion framework. As we demonstrate, MAS avoids common issues such as out-of-domain sampling and mode-collapse. https://guytevet.github.io/mas-page/

1. Introduction

3D motion generation is an increasingly popular field that has prominent applications in computer-animated films, video games, virtual reality, and more. One of the main bottlenecks of current approaches is reliance on 3D data, which is typically acquired by actors in motion capture studios or created by professional animation artists. Both forms of data acquisition are costly, not scalable, do not capture in-the-wild behavior, and leave entire motion domains under-explored.

Fortunately, the ubiquity of video cameras leads to countless high-quality recordings of a wide variety of motions. Naively, a possible way to leverage these videos for motion generation tasks is extracting 3D pose estimations and using them as training data. However, pose estimation methods are mostly trained using 3D data [19, 31], thus inheriting the mentioned data limitations. Some methods only require 2D data [5, 37], but suffer from noticeable artifacts and temporal inconsistencies.

Recently, Azadi et al. [2] and Zhang et al. [45] incorporated 3D motions estimated from images or videos into motion synthesis applications. The former used them to enrich an existing motion capture dataset and the latter as reference motions while learning a physically-based Rein-

forcement Learning policy. In both cases, the quality issues were bridged using strong priors (either high-quality 3D data or physical simulation), hence remaining limited to specific settings. Contrary to the pose estimation approaches, we focus on unconditional 3D motion generation from pure noise.

In this paper, we present Multi-view Ancestral Sampling (MAS), a diffusion-based 3D motion generation method, requiring only 2D motion data that can be acquired exclusively from videos. First, we learn a 2D motion diffusion model from a set of videos, then, we employ the MAS algorithm to effectively sample 3D motions from this learned model. Our method is based on *Ancestral Sampling* — the standard denoising loop used for sampling from a diffusion model. MAS extends this concept and generates a 3D motion by simultaneously denoising multiple 2D views describing it. At each diffusion denoising step, all views are triangulated into a single 3D motion and then projected back to each view. This ensures multi-view consistency throughout the denoising process, while adhering to the prior's predictions. We further encourage multi-view consistency by projecting a 3D noise to each view whenever sampling from a Gaussian distribution in the 2D ancestral sampling process.

We show that MAS generates diverse and realistic motions from the underlying 3D motion distribution using a 2D diffusion model that was exclusively trained on motions obtained from in-the-wild videos. Furthermore, we show that relying on ancestral sampling allows MAS to generate a 3D motion in a few seconds only, using a single standard GPU. MAS excels in scenarios where acquiring 3D motion capture data is impractical while video footage is abundant (See Figure 1). In such settings, we apply off-the-shelf 2D pose estimators to extract 2D motion sequences from video frames, and use them to train our diffusion prior. We demonstrate MAS in three domains: (1) professional basketball player motions extracted from common NBA match recordings, (2) horse motions extracted from equestrian contests, and (3) human-ball interactions extracted from rhythmic ball gymnastics performances (ball location is an additional parameter predicted by the model). These datasets demonstrate motion domains that were previously under-explored due to 3D data scarcity.

2. Related Work

3D Motion Synthesis. Multiple works explore 3D motion generation using moderate-scale 3D motion datasets such as HumanML3D [8], KIT-ML [24] Human3.6M [14] and HumanAct12 [7]. With this data, synthesis tasks were traditionally learned using Auto-Encoders or VAEs [18], [1, 8, 11, 23, 35]. Recently, Denoising Diffusion Models [33, 34] were introduced to this domain by MDM [36], MotionDiffuse [46], MoFusion [4], and FLAME [17]. Diffusion models were proven to have a better capacity to model the motion distribution of the data and provided opportunities for new generative tasks. Yet the main limitation of all the mentioned methods is their reliance on high-quality 3D motion capture datasets, which are hard to obtain and limited in domain and scale. In this context, SinMDM [27] enabled non-humanoid motion learning from a single animation; PriorMDM [29] and GMD [15] presented fine-tuning and inference time applications for motion tasks with few to none training samples, relying on a pre-trained MDM.

Monocular Pose Estimation. Monocular 3D pose estimation is a well-explored field [19, 30, 31, 43]. Its main challenge is the many ambiguities (e.g. self-occlusions and blurry motion) inherent to the problem. A parallel line of work is pose lifting from 2D to 3D. MotionBERT [48] demonstrates a supervised approach to the task. Some works offer to only use 2D data and learn in an unsupervised manner; Drover et al. [6] suggest training a 2D discriminator to distinguish between random projections of outputs of a 3D lifting network and the 2D data while optimizing the lifting network to deceive the discriminator; ElePose [37] train a normalizing-flows model on 2D poses and then use it to guide a 3D lifting network to generate 3D poses that upon projection have high probability w.r.t the normalizing-flows model. They add self-consistency and geometric losses and also predict the elevation angle of the lifted pose which is crucial for their success.

Animal 3D Shape Reconstruction. The recent MagicPony [41] estimates the pose of an animal given a single image by learning a per-category 3D shape template and per-instance skeleton articulations, trained to reconstruct a set of 2D images upon rendering. Yao et al. [42] suggest a method for improving the input images with occlusions/truncation via 2D diffusion. Then, they use a text-to-image diffusion model to guide 3D optimization process to obtain shapes and textures that are faithful to the input images.

Text to 3D Scene Generation. DreamFusion [25] and SJC [38], introduced guidance of 3D content creation using diffusion models trained on 2D data. Poole et al. [25] suggest SDS, a method for sampling from the diffusion model by minimizing a loss term that represents the distance between the model's distribution and the noised sample distribution. They suggest to harness SDS for 3D generation by repeatedly rendering a 3D representation (mostly NeRF [22] based) through a differentiable renderer, noising the resulting images using the forward diffusion, get a correction direction using the diffusion model, and then back-propagate gradients to update the 3D representation according to the predicted corrections. Although promising, their results are of relatively low quality and diversity and suffer from slow inference speed, overly saturated col-

ors, lack of 3D consistency, and heavy reliance on text conditioning. Follow-up works such as ProlificDreamer [40], HIFA [47], DreamTime [13], DDS [10] and NFSD [16] expose those weaknesses and suggest various methods to mitigate them. In a similar context, Instruct-NeRF2NeRF [9] edit a NeRF by gradually editing its source multi-view image dataset during training, using an image diffusion model. MVDream [32] train a diffusion model to generate multiple views of the same object using a 3D object dataset. They apply SDS optimization loop using the diffusion model to correct multiple views of the optimizing object at each iteration. This method and similar ones [12, 21, 28, 44] heavily rely on additional data such as 3D structure, depth or normals, which is not available in our setting.

Contrary to the SDS approach which is an optimization process, our MAS samples 3D motions from 2D diffusion models at inference. Hence it suggests a faster approach and avoids many of the SDS weaknesses by design (See Section 5).

3. Preliminary

Diffusion Models and Ancestral Sampling. Diffusion models are generative models that learn to gradually transform a predefined noise distribution into the data distribution. For the sake of simplicity, we consider the source distribution to be Gaussian. The forward diffusion process is defined by taking a data sample and gradually adding noise to it until we get a Gaussian distribution. The diffusion denoising model is then parameterized according to the reverse of this process, i.e. the model will sample a random Gaussian sample and gradually denoise it until getting a valid sample.

Formally, the forward process is defined by sampling a data sample $x_0 \sim q(x_0)$ and for t in $1, ..., T$, sampling $x_t \sim q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$, until getting to x_T, which has a gaussian distribution $x_T \sim q(x_T) = \mathcal{N}(x_T; 0, I)$.

The reverse process, also called **ancestral sampling**, is defined by sampling a random gaussian noise $x_T \sim p_\phi(x_T) = \mathcal{N}(x_T; 0, I)$ and then for t in $T, t-1, ..., 1$, sampling $\hat{x}_{t-1} \sim p_\phi(\hat{x}_{t-1}|x_t)$, until getting to $\hat{x}_0$, which should ideally approximate the data distribution. The model posterior $p_\phi(x_{t-1}|x_t)$ is parameterized by a network $\mu_\phi(x_t, t)$:

$$p_\phi(x_{t-1}|x_t) = q(x_{t-1}|x_t, x_0 = \mu_\phi(x_t; t)) = \mathcal{N}\left(x_{t-1}; \mu_\phi(x_t, t), \sigma_t^2 I\right)$$

i.e. the new network predicts a mean denoising direction from x_t which is then used for sampling x_{t-1} from the posterior distribution derived from the forward process. μ_ϕ is further parameterized by a network ϵ_ϕ that aims to predict the noise embedded in x_t:

$$\mu_\phi(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\phi(x_t, t)\right)$$

Now, when optimizing the usual variational bound on negative log-likelihood, it simplifies to,

$$\mathcal{L}(\phi) = \mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(t)\|\epsilon_\phi(\alpha_t x_0 + \sigma_t\epsilon; t) - \epsilon\|_2^2\right]$$

which is used as the training loss. We approximate this loss by sampling t, ϵ, x_0 from their corresponding distributions and calculating the loss term. Note that when adding text-conditioning to the model, it is denoted by $p_\phi(x|y)$ where y is the text prompt.

Data Representation. A motion sequence is defined on top of a character skeleton with J joints. A single character pose is achieved by placing each joint in space. Varying the character pose over time constructs a motion sequence. Hence, we denote a 3D motion sequence, $X \in \mathbb{R}^{L\times J\times 3}$, with L frames by the xyz location of each joint at each frame. Note that this representation is not explicitly force fixed bone length. Instead, our algorithm will do so implicitly. Additionally, This formulation allows us to model additional moving objects in the scene (e.g. a ball or a box) using auxiliary joints to describe their location.

Considering the pinhole camera model[1], we define a camera-view $v = (R_v, \tau_v, f_v)$ by its rotation matrix $R_v \in \mathbb{R}^{3\times 3}$, translation vector $\tau_v \in \mathbb{R}^3$ and the focal length f_v given in meters. Then, a 2D motion, $x^v = P(X, v) \in \mathbb{R}^{L\times J\times 2}$, from camera-view v, is defined as the perspective projection P of X to v such that each joint at each frame is represented with its uv coordinates of the camera space.

In order to drive 3D rigged characters (as presented in the figures of this paper) we retrieve 3D joint angles from the predicted 3D joint positions of X using SMPLify [3] optimization for human characters, and Inverse-Kinematics optimization for the non-humanoid characters (i.e. horses).

4. Method

Our goal is to generate 3D motion sequences using a diffusion model trained on monocular 2D motions. This would enable 3D motion generation in the absence of high-quality 3D data, by leveraging the ubiquity of monocular videos describing those scenes. To this end, we introduce Multi-view Ancestral Sampling (MAS), a method that simultaneously generates multiple views of a 3D motion via ancestral sampling. MAS maintains consistency between the 2D motions in all views at each denoising step to construct a coherent 3D motion. A single MAS step is illustrated in Figure 3.

In our experiments we first extract 2D pose estimations from in-the-wild videos and use them to train a 2D diffusion

[1] https://en.wikipedia.org/wiki/3D_projection#Perspective_projection

model $\hat{x}_0 = G_{2D}(x_t)$, that predicts the clean 2D motion, $\hat{x}_0$ at each denoising step (See Figure 2).

MAS then uses the diffusion model to simultaneously apply an ancestral sampling loop on multiple 2D motions, which represent views of the same 3D motion from V different camera angles. At each denoising step t, we get a set of noisy views $x_t^{1:V}$ as input and predict clean samples $\hat{x}_0^{1:V} = G_{2D}(x_t^{1:V})$. Then, the *Consistency Block* is applied in two steps: (1) Triangulation: find a 3D motion X that follows all views as closely as possible. (2) Reprojection: project the resulting 3D motion to each view, getting $\tilde{x}_0^{1:V}$, which we can think of as a multiview-consistent version of the predicted motions. Finally, we can sample the next step $x_{t-1}^{1:V}$ from the backward posterior $x_{t-1}^{1:V} \sim q\left(x_{t-1}|x_t, \tilde{x}_0^{1:V}\right)$ just like the original ancestral sampling algorithm. Repeating this denoising process up to $t = 0$ yields multiple views of the same 3D motion. Finally, we triangulate the resulting 2D motions to create a 3D motion, which is returned as the final output. This sampling process is detailed in the Supp. The remainder of this section describes the monocular data collection and diffusion pretraining (4.1), followed by a full description of MAS building blocks (4.2).

4.1. Preparations

Data Collection. We collect videos from various sources — NBA videos, horse jumping contests, and rhythmic gymnastics contests. We then apply multi-person and object tracking using off-the-shelf models to extract bounding boxes. Subsequently, we use other off-the-shelf models for 2D pose estimation to get 2D motions. Implementation details are in Section 6. We build on the fact that 2D pose estimation is a well-explored topic, with large-scale datasets that can be easily scaled as manual annotations are much easier to obtain compared to 3D annotation which usually requires a motion capture studio.

2D Diffusion Model Training. We follow Tevet et al. [36] and train the unconditioned version of the Motion Diffusion Model (MDM) with a transformer encoder backbone for each of the datasets separately. We boost the sampling of MDM by a factor of 10 by learning 100 diffusion steps instead of the original 1000.

4.2. Multi-view Ancestral Sampling

We would like to construct a way to sample a 3D motion using a model that generates 2D samples. First, we observe that a 3D motion is uniquely defined by 2D views of it from multiple angles. Second, we assume that our collected dataset includes a variety of motions, from multiple viewpoints, and deduce that our 2D diffusion model can generalize for generating multiple views of the same 3D motion, for a wide variety of 3D motions. Thus, we aim to generate multiple 2D motions that represent multiple views of the same 3D motion, from a set of different view-points.

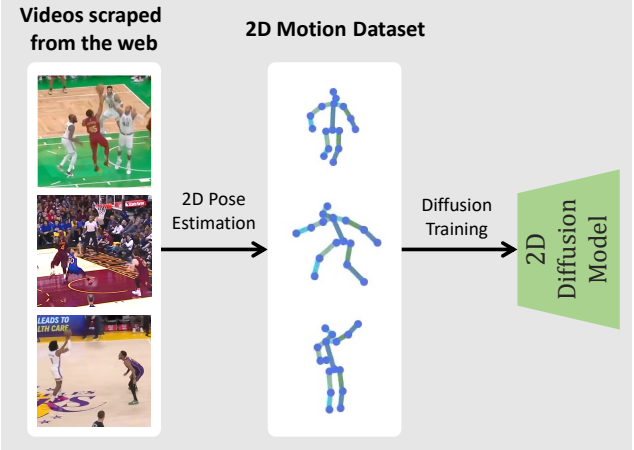

Figure 2. **Preparations.** The motion diffusion model used for MAS is trained on 2D motion estimations of videos scraped from the web.

Ancestral Sampling for 3D generation. As described in Section 3, diffusion models are designed to be sampled using gradual denoising, following the ancestral sampling scheme. Hence, we design MAS to generate multiple 2D motions via ancestral sampling, while guiding all views to be multiview-consistent. Formally, we take a set of V views, distributed evenly around the motion subject, with elevation angle distribution heuristically picked for each dataset. Then, for a each view v we initialize x_T^v with noise, and for $t = T, ..., 1$ transform x_t^v to x_{t-1}^v until getting a valid 2D motion x_0^v for each view. We choose to generate all views concurrently, keeping all views in the same diffusion timestep throughout the process.

In every denoising step we receive $x_t^{1:V} = \left(x_t^1, ..., x_t^V\right)$. We derive the clean motion predictions by applying the diffusion model in each view $\hat{x}_0^v := \frac{x_t^v - \sqrt{1-\bar{\alpha}_t}\epsilon_\phi(x_t^v)}{\sqrt{\bar{\alpha}_t}}$, getting $\hat{x}_0^{1:V} = \left(\hat{x}_0^1, ..., \hat{x}_0^V\right)$. We apply our multi-view Consistency Block to find multi-view consistent motions $\tilde{x}_0^{1:V}$ that approximate the predicted motions $\hat{x}_0^{1:V}$. We then use the resulting motions $\tilde{x}_0^{1:V}$ as the denoising direction by sampling x_{t-1}^v from $q\left(x_{t-1}^v|x_t^v, x_0 = \tilde{x}_0^v\right)$, and outputting $x_{t-1}^{1:V} = \left(x_{t-1}^1, ..., x_{t-1}^V\right)$.

MAS can be extended to support dynamic camera-view along sampling instead of fixed ones as detailed in The Supp. Since this is not empirically helpful for our application, we leave it out of our scope.

Multi-view Consistency Block As mentioned, the purpose of this block is to transform multiview motions $\hat{x}_0^{1:V}$ into multiview-consistent motions $\tilde{x}_0^{1:V}$ that are as similar as possible. We achieve this by finding a 3D motion X that when projected to all views, it resembles the multiview mo-

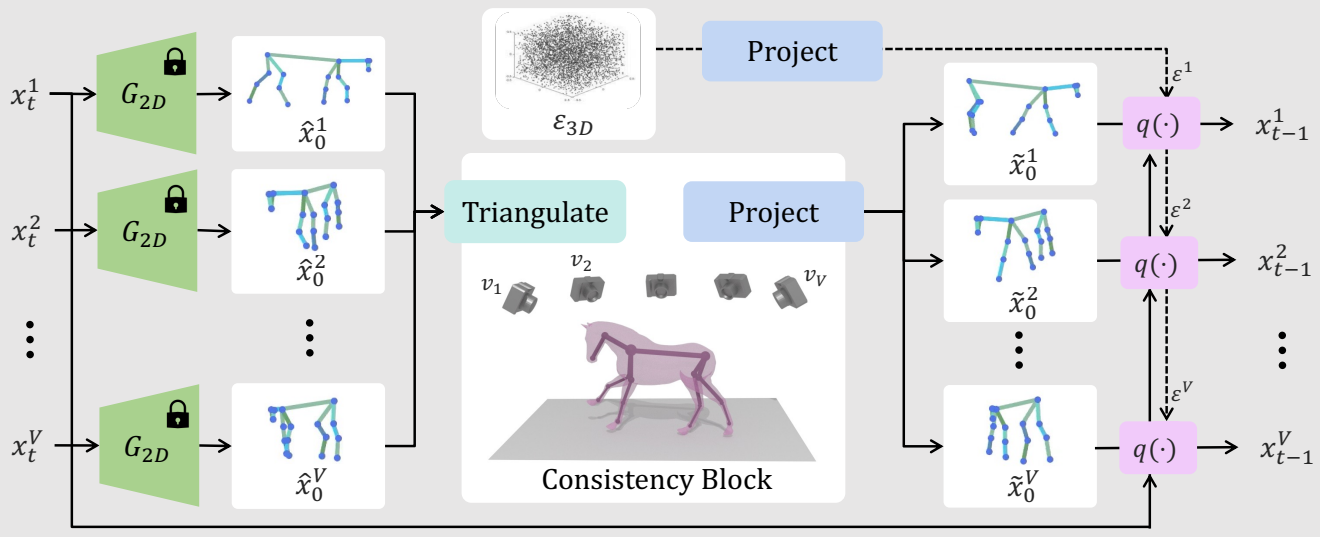

Figure 3. The figure illustrates an overview of MAS, showing a multi-view denoising step from the 2D sample collection $x_t^{1:V}$ to $x_{t-1}^{1:V}$, corresponding to camera views $v_{1:V}$. Denoising is performed by a fixed 2D motion diffusion model G_{2D}. At each such iteration, our *Consistency Block* triangulates the motion predictions $\hat{x}_0^{1:V}$ into a single 3D sequence and projects it back onto each view ($\tilde{x}_0^{1:V}$). To encourage consistency in the model's predictions, we sample 3D noise, ϵ_{3D} and project it to the 2D noise ϵ^v for each view. Finally, we sample $x_{t-1}^{1:V}$ from $q\left(x_{t-1}^{1:V}|x_t^{1:V}, \tilde{x_0}^{1:V}\right)$.

tions $\hat{x}_0^{1:V}$ via *Triangulation*. We then return projections of X to each view $\tilde{x}_0^{1:V} = (P(X,1), ..., P(X,V))$, as the multiview-consistent motions. Since the denoising process is gradual, the model's predictions are approximately multiview-consistent so the consistency block only makes small corrections.

Triangulation. We calculate X via optimization to minimize the difference between projections of X to all views and the multiview motion predictions $\hat{x}_0^{1:V}$:

$$X = \arg\min_{X'} \|P(X', 1{:}V) - \hat{x}_0^{1:V}\|_2^2 = \arg\min_{X'} \sum_{v=1}^{V} \|P(X', v) - \hat{x}_0^v\|_2^2$$

For faster convergence, we initialize X with the optimized results from the previous sampling step. This way the process can also be thought of progressively refining X but we wish to emphasize that the focus remains the ancestral sampling in the 2D views.

3D Noise. When triangulating the 2D motions $\hat{x}_0^{1:V}$, we would like them to be as close to being multiview-consistent as possible. A critical observation is that for our model to generate multiview-consistent motions we would like to pass it multiview-consistent noised motions. To this end, we design a new noise sampling mechanism that will (1) keep Gaussian distribution for each view, and (2) maintain multiview-consistency.

We start by sampling 3D noise $\varepsilon_{3d} \sim \mathcal{N}(0, I)$ ($\varepsilon_{3d} \in \mathbb{R}^{L \times J \times 3}$). Projecting this noise to each view using perspective projection will result in a distribution that is not Gaussian. Hence, we instead use orthographic projection, which preserves Gaussian distribution for each view (See Supp.), and can differ from perspective projection by at most $O(1/(d-1))$, where d is the distance between the camera and the subject's center and assuming the subject is normalized to be bounded in a sphere with radius 1 (See Supp.) We then use the resulting distribution for sampling the initial noise x_T and when sampling $x_{t-1} \sim q(x_{t-1}|x_t, x_0 = P(X))$ which significantly improves the quality and diversity of our results (see table 3).

5. Method Discussion

In this section, we discuss the properties of MAS, contextualizing it within the landscape of recent advancements in the text-to-3D domain.

Ancestral sampling. MAS is built upon the ancestral sampling process. This means that the model is used in its intended way over in-domain samples. This is in contrast to SDS-based methods [25] which employ a sampling scheme that uses the forward diffusion to noise images rendered from a 3D representation that is only partially optimized. This can lead to out-of-distribution samples, particularly when using smaller timesteps where the model expects motions that are close to being real. This phenomenon is also addressed by [38] and [13], who suggest heuristics

Dataset Name	Subject	#Samples	Length Range	Average Length	FPS	In-the-wild videos
Human3.6M [14]	Humans	300	42s-240s	104s	25	✗
NBA videos	Humans	$60K$	4s-16s	6s	30	✓
Horse jumping contests	Horses	$2K$	3s-40s	7s	20	✓
Rhythmic ball gymnastics	Humans + Ball	500	10s-120s	81s	20	✓

Table 1. **2D Datasets.** Details of the 2D motion datasets used for our experiments. The last three are newly collected in-the-wild datasets which we made available at https://guytevet.github.io/mas-page/.

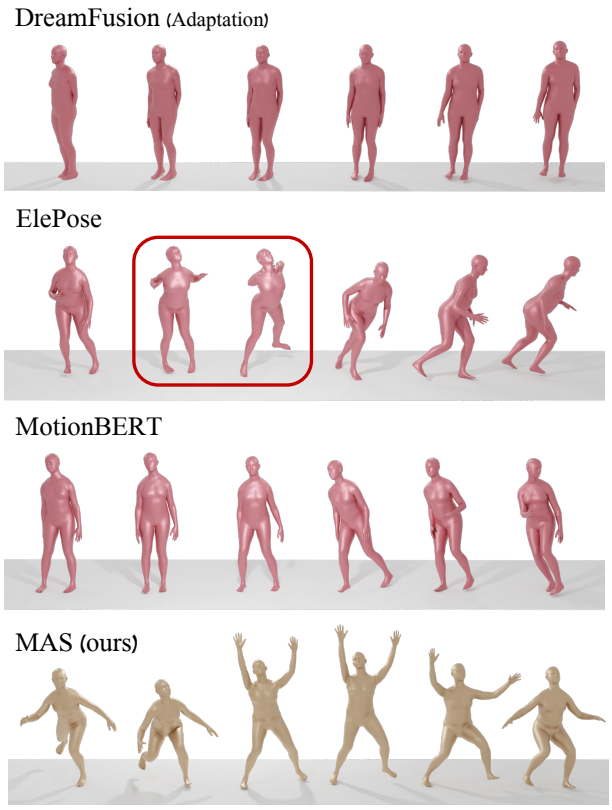

Figure 4. Generated motions by MAS compared to ElePose [37], MotionBert [48], and an adaptation of DreamFusion [25] to unconditioned motion generation. We observe that MotionBert and DreamFusion produce dull motions with limited movement and ElePose predictions are jittery and often include invalid poses (Red rectangles).

to alleviate the out-of-distribution problem but do not fundamentally solve it. Furthermore, most SDS-based methods sample x_t independently in each iteration, which may lead to a high variance in the correction signal. Contrarily, using ancestral sampling has, by definition, a large correlation between x_t and x_{t-1}, which leads to a more stable process and expressive results. Since MAS is sampling-based, it naturally models the diversity of the distribution, while optimization-based methods often experience mode-collapse or divergence, as addressed by [25]. It is worth noting that SDS is a clever design for cases where ancestral sampling cannot be used.

Multi-view stability. MAS simultaneously samples multiple views that share the same timestep at each denoising step. SDS-based methods typically use a single view in each optimization step, forcing them to make concessions such as small and partial corrections to prevent ruining the 3D object from other views. This also leads to a state where it is unknown which timestep to choose, since only partial denoising steps were applied (also shown by [13]). MAS avoids such problems since the multiview denoising steps are applied simultaneously. It allows us to apply full optimization during the triangulation process. Hence, by the end of the i'th iteration, each view follows the model's distribution at timestep $T - i$. This alleviates the need for timestep scheduling and avoids out-of-distribution samples.

3D noise consistency. MAS's usage of a multiview-consistent noise distribution, critically boosts multiview-consistency in the model's predictions and greatly benefits the quality and diversity of the generated motions. SDS-based methods sample uncorrelated noise in different views, which leads to inconsistent corrections, that can result in a lack of 3D consistency, slower convergence or even divergence.

6. Experiments

6.1. Data Collection

In order to demonstrate the merits of our method, we apply MAS on three different 2D motion datasets. Each dataset addresses a different motion aspect that is under-represented in existing 3D motion datasets (See Table 1). (1) The NBA players' performance dataset demonstrates motion generation in domains of human motions that are poorly covered by existing datasets. (2) The horse show-jumping contests dataset shows generation in a domain that has almost no 3D data at all and has a completely different topology. Finally, (3) the rhythmic-ball gymnastics dataset shows that our method opens the possibility to model interactions with dynamic objects. All datasets include motions from diverse views, which is crucial for the success of our method. We detail the data collection process in the Supp.

In addition, we evaluate MAS on the 3D motion dataset, Human3.6M [14], by projecting the motions to random 2D

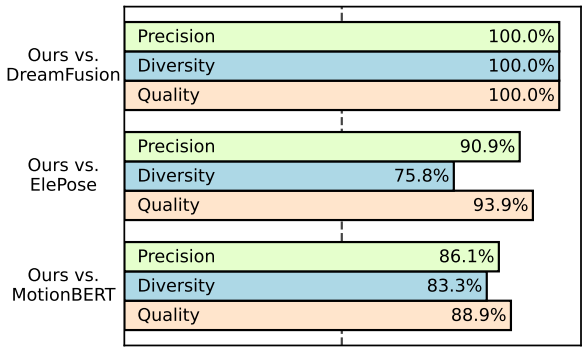

Figure 5. **NBA Dataset User study.** We asked 22 unique users to compare 15 randomly generated motions by each of the models to MAS generations in 3 aspects - *precision* (i.e. what samples best depict Basketball moves), Overall *Quality* and *Diversity*. The dashed line marks 50%. MAS outperforms the lifting methods and the DreamFusion adaptation.

cameras.

All motions are represented as $x \in \mathbb{R}^{L\times J\times 2}$ as was detailed in Section 3, where NBA is using the AlphaPose body model with 16 joint, horses represented according to APT-36K with 17 joints and the gymnastics dataset is represented with the COCO body model [20] with 17 joints plus additional joint for the ball. All 2D pose predictions are accompanied by confidence predictions per joint per frame which are used in the diffusion training process.

6.2. Implementation Details

Our 2D diffusion model is based on MDM [36], and composed of a transformer encoder with 6 attention layers of 4 heads and a latent dimension of 512. This backbone supports motions with variable length in both training and sampling, which makes MAS support it as well. To mitigate some of the pose prediction errors, we mask low-confidence joint predictions from the training loss. We used an ADAM optimizer with 10^{-4} lr for training and cosine noise scheduling. We learn 100 diffusion steps instead of 1000 which accelerate MAS 10-fold without compromising the quality of the results. We observe that MAS performs similarly for any $V \geq 3$ and report 5 camera views across all of our experiments. The camera views $v_{1:V}$ are fixed through sampling, surrounding the character and sharing the same elevation angle, with azimuth angles evenly spread around $[0, 2\pi]$. Generating a 3D sample with MAS takes less than 10 seconds on a single NVIDIA GeForce RTX 2080 Ti. Performance details can be found in the Supp.

6.3. Evaluation

Here we explore the quality of the 3D motions generated by our method. Our experiments are conducted on the NBA dataset to allow comparison with existing methods, which mostly explore human motion. Usually, we would compare the generated motions to motions sampled from the dataset. In our case, we do not have 3D data so we must introduce a new way to evaluate the 3D generated motions. For that sake, we rely on the assumption that a 3D motion is of high-quality if and only if all 2D views of it are of high-quality. Consequently, we suggest taking random projections of the 3D motions and comparing them with our 2D data. More specifically, we generate a set of 3D motions, with lengths sampled from the data distribution, then sample a single angle for every motion with yaw drawn from $\mathcal{U}[0, 2\pi]$ and a constant pitch angle fitted for each dataset. We project the 3D motion to the sampled angle using perspective projection, from a constant distance (also fitted for each dataset) and get a set of 2D motions.

Finally, we follow common evaluation metrics [26, 36] used for assessing unconditional generative models: *FID* measures Fréchet inception distance between the generated data distribution and the test data distribution; *Diversity* measures the variance of generated motion in latent space; *Precision* measures the portion of the generated data that is covered by the test data; *Recall* measures the portion of the test data distribution that is covered by the measured distribution. These metrics are predominantly calculated in latent space. Hence, we train a VAE-based evaluator for each dataset. We evaluate over $1K$ random samples and repeat the process 10 times to calculate the average value and confidence intervals. Table 3 shows that MAS results are comparable to the diffusion model in use, which marks a performance upper bound in 2D. We show that the addition of the multiview-consistent noise is crucial to the success of our method and prevents mode collapse. A thorough ablation study for the number of views, camera distance, and number of diffusion steps can be found in the Supp.

We evaluate an adaptation of DreamFusion [25] to the unconditioned motion generation domain and show that it performs poorly. This is carried out by initializing a random 3D motion and then performing 200 SDS iterations using the same diffusion model we used for MAS. Each iteration is comprised of: (1) Projecting the 3D motion to some random view (view distribution is the same as in MAS).(2) Noising the resulting 2D motion to some diffusion timestep $t \sim \mathcal{U}[1, T]$. (3) Letting our diffusion model predict a cleaner version of the noised motion. (4) Updating the 3D motion to fit the predicted motion in the sampled view using a single optimization step. The implementation of this adaptation can be found in our published code.

We also experimented with higher iteration numbers than 200 and techniques such as timestep scheduling and optimization tuning but saw no significant improvement.

We compare our method with off-the-shelf SOTA methods for supervised pose lifting - MotionBERT [48] - and

	FID↓		Diversity→		Precision↑		Recall↑	
View Angles	All	Side	All	Side	All	Side	All	Side
Ground Truth	$1.05^{\pm.02}$		$8.97^{\pm.05}$		$0.73^{\pm.01}$		$0.73^{\pm.01}$	
ElePose [2021]	$10.76^{\pm.45}$	$18.28^{\pm.33}$	$9.72^{\pm.05}$	$\mathbf{8.98^{\pm.06}}$	$0.28^{\pm.02}$	$0.26^{\pm.02}$	$0.58^{\pm.03}$	$0.17^{\pm.01}$
MotionBert [2023]	$30.22^{\pm.26}$	$36.89^{\pm.40}$	$9.57^{\pm.09}$	$8.67^{\pm.08}$	$0.04^{\pm 4e-03}$	$0.03^{\pm.01}$	$0.34^{\pm.04}$	$0.15^{\pm.04}$
MAS (Ours)	$\mathbf{5.38^{\pm.06}}$		$9.47^{\pm.06}$		$\mathbf{0.50^{\pm.01}}$		$\mathbf{0.60^{\pm.01}}$	

Table 2. **Comparison with pose lifting on NBA dataset.** MAS outperforms state-of-the-art unsupervised lifting methods. Furthermore, lifting methods experience a drop in recall when evaluated from the side view ($\mathcal{U}\left(\frac{\pi}{4}, \frac{3\pi}{4}\right)$), while MAS does not suffer from this limitation as it is a generative approach, and not lifting-based. '→' means results are better when the value is closer to the real distribution (8.97 for Diversity); **bold** marks best results.

	FID↓	Diversity→	Precision↑	Recall↑
Ground Truth	$1.05^{\pm.02}$	$8.97^{\pm.05}$	$0.73^{\pm.01}$	$0.73^{\pm.01}$
2D Diffusion Model	$5.23^{\pm.13}$	$9.70^{\pm.08}$	$0.44^{\pm.02}$	$0.78^{\pm.01}$
MAS (Ours)	$\mathbf{5.38^{\pm.06}}$	$\mathbf{9.47^{\pm.06}}$	$\mathbf{0.50^{\pm.01}}$	$0.60^{\pm.01}$
with 2 views (120°)	$6.87^{\pm.14}$	$9.99^{\pm.06}$	$0.35^{\pm.01}$	$\mathbf{0.80^{\pm.01}}$
- 3d noise	$17.40^{\pm.12}$	$6.67^{\pm.07}$	$0.93^{\pm.01}$	$0.01^{\pm 2.6e-03}$
DreamFusion [2022]	$66.38^{\pm 1.24}$	$8.25^{\pm.16}$	$0.33^{\pm.08}$	$0.17^{\pm.13}$

Table 3. **Ablations.** We compare MAS to an adaptation of DreamFusion [25] to the unconditional motion generation domain. Our evaluation measures the quality of 2D projections of the 3D generated motions. Our ablations show that MAS performs best with as few as 5 views (ours), and 3D noise is crucial for preventing mode collapse. gray indicates mode-collapse (Recall$< 10\%$), **bold** marks the best results otherwise. '→' means results are better when the value is closer to the real (train data) distribution.

unsupervised pose lifting - ElePose [37]. Although these methods are not generative per-se, we consider lifted motions from 2D motions sampled from the training data as generated samples. As Elepose only requires 2D data, we train it on our NBA dataset and adjust the geometric priors to our data. MotionBert was trained on Human3.6M [14] dataset and some in-the-wild videos, so it is applied in a zero-shot setting. Table 2 shows that MAS outperforms both lifting methods.

Since we sample a uniform angle around the lifted motions, we often project them to views that are similar to the lifted view. This results in a motion that resembles the lifted motion, which was sampled from the train data, thus boosting performance. We show that when evaluating from the side view (angle $\sim \mathcal{U}\left(\frac{\pi}{4}, \frac{3\pi}{4}\right)$ relative to the lifting angle)), the lifting methods experience a clear degradation in performance. MAS is unaffected as it is a generative approach and has no "side" view. Repeating this experiment with the 3D dataset Human3.6M, randomly projected into 2D cameras shows that MAS is on par with the side-view performance of MotionBERT, and ElePose. More details in the Supp. Figure 4 demonstrates the quality of MAS compared to DreamFusion, MotionBERT, and ElePose. Figure 5 presents a user study conducted with 22 participants comparing 15 randomly generated 3D motions by each of the models. An example screenshot from the study can be found in the Supp.

7. Conclusions

In this paper, we introduced MAS, a generative method designed for 3D motion synthesis using 2D data. We showed that high-quality 3D motions can be sampled from a diffusion model trained on 2D data only. The essence of our method lies in its utilization of a multiview diffusion ancestral sampling process, where each denoising step contributes to forging a coherent 3D motion sequence.

Our experiments show that MAS excels with in-the-wild videos, enabling it to produce motions that are otherwise exceedingly challenging to obtain through conventional means.

Our method could also be employed in additional domains such as multi-person interactions, hand and face motions, complex object manipulations and with recent developments in tracking of "any" object [39], we wish to push the boundaries of data even further.

Our method does experience some failure cases: The character occasionally folds into itself when changing direction, and the character sometimes changes its scale throughout the motion. MAS also inherits the limitations of the 2D data it is using and thus cannot naively predict global position, or apply textual control. We leave extending the data acquisition pipeline to support such features to future work. It is also worth noting that our method requires 2D data that captures a variety of views of similar motions. Finally, we hope the insights introduced in this paper can also be utilized in the text-to-3D field and other applications.

Acknowledgements

We thank Elad Richardson, Inbar Gat, and Matan Cohen for thoroughly reviewing our early drafts. We thank Sigal Raab, Oren Katzir, and Or Patashnik for the fruitful discussions. This research was supported in part by the Israel Science Foundation (grants no. 2492/20 and 3441/21), Len Blavatnik and the Blavatnik family foundation, and The Tel Aviv University Innovation Laboratories (TILabs). This work was supported by the Yandex Initiative in Machine Learning.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 2

[2] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation, 2023. 1

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 3

[4] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 2

[5] Yicheng Deng, Cheng Sun, Jiahui Zhu, and Yongqi Sun. Sv-mac: Unsupervised 3d human pose estimation from a single image with single-view-multi-angle consistency, 2022. 1

[6] Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone?, 2018. 2

[7] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 2

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 2

[9] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[10] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. *arXiv preprint arXiv:2304.07090*, 2023. 3

[11] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[12] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation, 2023. 3

[13] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation, 2023. 3, 5, 6

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 6, 8

[15] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Gmd: Controllable human motion synthesis via guided diffusion models. *arXiv preprint arXiv:2305.12577*, 2023. 2

[16] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 3

[17] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 7

[21] 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, 2023. 3

[22] 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, 2020. 2

[23] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[24] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2

[25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5, 6, 7, 8

[26] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. *arXiv preprint arXiv:2206.08010*, 2022. 7

[27] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 2

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

[29] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

[30] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation, 2023. 2

[31] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation, 2023. 1, 2

[32] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2023. 3

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[34] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2

[35] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 2

[36] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4, 7

[37] Bastian Wandt, James J. Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses, 2021. 1, 2, 6, 8

[38] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. 2, 5

[39] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 8

[40] 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, 2023. 3

[41] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild, 2023. 2

[42] Chun-Han Yao, Amit Raj, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Artic3d: Learning robust articulated 3d shapes from noisy web image collections, 2023. 2

[43] Bruce X. B. Yu, Zhi Zhang, Yongxu Liu, Sheng hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video, 2023. 2

[44] Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation, 2023. 3

[45] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph.*, 2023. 1

[46] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[47] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance, 2023. 3

[48] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations, 2023. 2, 6, 7, 8