HUMAN MOTION DIFFUSION MODEL

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir,
Daniel Cohen-Or and Amit H. Bermano
Tel Aviv University, Israel
guytevet@mail.tau.ac.il

ABSTRACT

Natural and expressive human motion generation is the holy grail of computer animation. It is a challenging task, due to the diversity of possible motion, human perceptual sensitivity to it, and the difficulty of accurately describing it. Therefore, current generative solutions are either low-quality or limited in expressiveness. Diffusion models, which have already shown remarkable generative capabilities in other domains, are promising candidates for human motion due to their many-to-many nature, but they tend to be resource hungry and hard to control. In this paper, we introduce Motion Diffusion Model (MDM), a carefully adapted classifier-free diffusion-based generative model for the human motion domain. MDM is transformer-based, combining insights from motion generation literature. A notable design-choice is the prediction of the sample, rather than the noise, in each diffusion step. This facilitates the use of established geometric losses on the locations and velocities of the motion, such as the foot contact loss. As we demonstrate, MDM is a generic approach, enabling different modes of conditioning, and different generation tasks. We show that our model is trained with lightweight resources and yet achieves state-of-the-art results on leading benchmarks for text-to-motion and action-to-motion[1]. https://guytevet.github.io/mdm-page/.

1 INTRODUCTION

Human motion generation is a fundamental task in computer animation, with applications spanning from gaming to robotics. It is a challenging field, due to several reasons, including the vast span of possible motions, and the difficulty and cost of acquiring high quality data. For the recently emerging text-to-motion setting, where motion is generated from natural language, another inherent problem is data labeling. For example, the label "kick" could refer to a soccer kick, as well as a Karate one. At the same time, given a specific kick there are many ways to describe it, from how it is performed to the emotions it conveys, constituting a many-to-many problem. Current approaches have shown success in the field, demonstrating plausible mapping from text to motion (Petrovich et al., 2022; Tevet et al., 2022; Ahuja & Morency, 2019). All these approaches, however, still limit the learned distribution since they mainly employ auto-encoders or VAEs (Kingma & Welling, 2013) (implying a one-to-one mapping or a normal latent distribution respectively). In this aspect, diffusion models are a better candidate for human motion generation, as they are free from assumptions on the target distribution, and are known for expressing well the many-to-many distribution matching problem we have described.

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020; Ho et al., 2020) are a generative approach that is gaining significant attention in the computer vision and graphics community. When trained for conditioned generation, recent diffusion models (Ramesh et al., 2022; Saharia et al., 2022b) have shown breakthroughs in terms of image quality and semantics. The competence of these models have also been shown for other domains, including videos (Ho et al., 2022), and 3D point clouds (Luo & Hu, 2021). The problem with such models, however, is that they are notoriously resource demanding and challenging to control.

In this paper, we introduce Motion Diffusion Model (MDM) — a carefully adapted diffusion based generative model for the human motion domain. Being diffusion-based, MDM gains from the na-

[1]Our code can be found at https://github.com/GuyTevet/motion-diffusion-model.

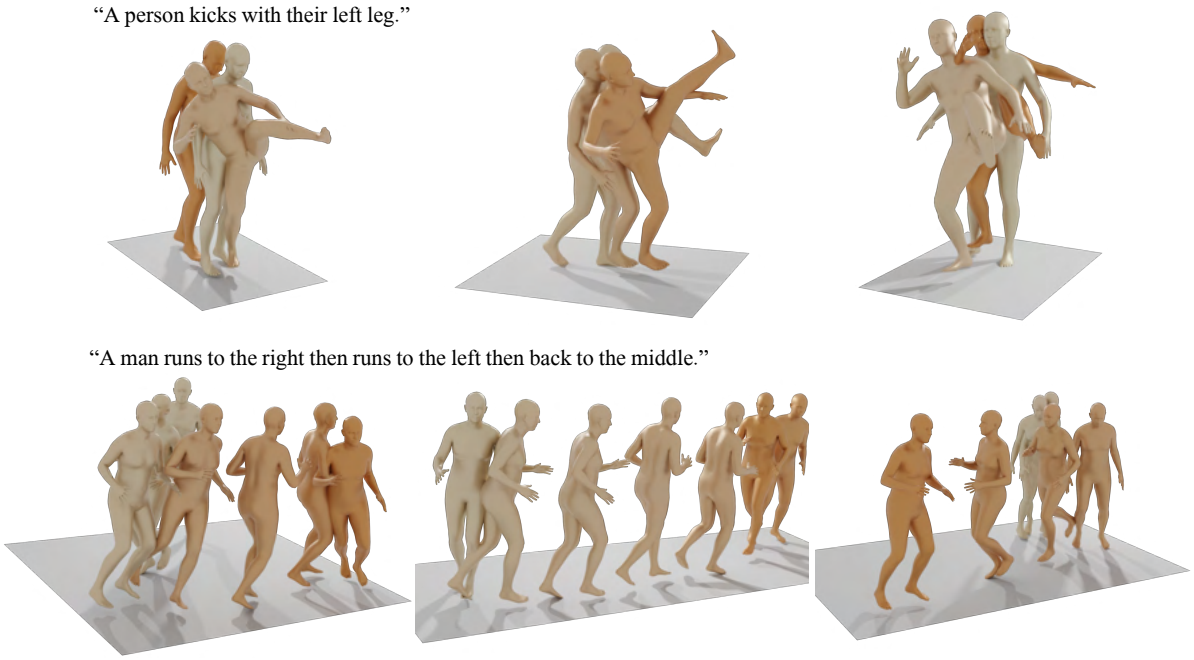


Figure 1: Our Motion Diffusion Model (MDM) reflects the many-to-many nature of text-to-motion mapping by generating diverse motions given a text prompt. Our custom architecture and geometric losses help yielding high-quality motion. Darker color indicates later frames in the sequence.

tive aforementioned many-to-many expression of the domain, as evidenced by the resulting motion quality and diversity (Figure 1). In addition, MDM combines insights already well established in the motion generation domain, helping it be significantly more lightweight and controllable.

First, instead of the ubiquitous U-net (Ronneberger et al., 2015) backbone, MDM is transformer-based. As we demonstrate, our architecture (Figure 2) is lightweight and better fits the temporal and non-spatial nature of motion data (represented as a collection of joints). A large volume of motion generation research is devoted to learning using geometric losses (Kocabas et al., 2020; Harvey et al., 2020; Aberman et al., 2020). Some, for example, regulate the velocity of the motion (Petrovich et al., 2021) to prevent jitter, or specifically consider foot sliding using dedicated terms (Shi et al., 2020). Consistently with these works, we show that applying geometric losses in the diffusion setting improves generation.

The MDM framework has a generic design enabling different forms of conditioning. We showcase three tasks: text-to-motion, action-to-motion, and unconditioned generation. We train the model in a classifier-free manner (Ho & Salimans, 2022), which enables trading-off diversity to fidelity, and sampling both conditionally and unconditionally from the same model. In the text-to-motion task, our model generates coherent motions (Figure 1) that achieve state-of-the-art results on the HumanML3D (Guo et al., 2022a) and KIT (Plappert et al., 2016) benchmarks. Moreover, our user study shows that human evaluators prefer our generated motions over real motions 42% of the time (Figure 4(a)). In action-to-motion, MDM outperforms the state-of-the-art (Guo et al., 2020; Petrovich et al., 2021), even though they were specifically designed for this task, on the common HumanAct12 (Guo et al., 2020) and UESTC (Ji et al., 2018) benchmarks.

Lastly, we also demonstrate completion and editing. By adapting diffusion image-inpainting (Song et al., 2020b; Saharia et al., 2022a), we set a motion prefix and suffix, and use our model to fill in the gap. Doing so under a textual condition guides MDM to fill the gap with a specific motion that still maintains the semantics of the original input. By performing inpainting in the joints space rather than temporally, we also demonstrate the semantic editing of specific body parts, without changing the others (Figure 3).

Overall, we introduce Motion Diffusion Model, a motion framework that achieves state-of-the-art quality in several motion generation tasks, while requiring only about three days of training on a

single mid-range GPU. It supports geometric losses, which are non trivial to the diffusion setting, but are crucial to the motion domain, and offers the combination of state-of-the-art generative power with well thought-out domain knowledge.

2 Related Work

2.1 Human Motion Generation

Neural motion generation, learned from motion capture data, can be conditioned by any signal that describes the motion. Many works use parts of the motion itself for guidance. Some predict motion from its prefix poses (Fragkiadaki et al., 2015; Martinez et al., 2017; Hernandez et al., 2019; Guo et al., 2022b). Others (Harvey & Pal, 2018; Kaufmann et al., 2020; Harvey et al., 2020; Duan et al., 2021) solve in-betweening and super-resolution tasks using bi-directional GRU (Cho et al., 2014) and Transformer (Vaswani et al., 2017) architectures. Holden et al. (2016) use auto-encoder to learn motion latent representation, then utilize it to edit and control motion with spatial constraints such as root trajectory and bone lengths. Motion can be controlled with a high-level guidance given from action class (Guo et al., 2020; Petrovich et al., 2021; Cervantes et al., 2022), audio (Li et al., 2021; Aristidou et al., 2022) and natural language (Ahuja & Morency, 2019; Petrovich et al., 2022). In most cases authors suggests a dedicated approach to map each conditioning domain into motion.

In recent years, the leading approach for the *Text-to-Motion* task is to learn a shared latent space for language and motion. JL2P (Ahuja & Morency, 2019) learns the KIT motion-language dataset (Plappert et al., 2016) with an auto-encoder, limiting one-to-one mapping from text to motion. TEMOS (Petrovich et al., 2022) and T2M (Guo et al., 2022a) suggest using a VAE (Kingma & Welling, 2013) to map a text prompt into a normal distribution in latent space. Recently, MotionCLIP (Tevet et al., 2022) leverages the shared text-image latent space learned by CLIP (Radford et al., 2021) to expand text-to-motion out of the data limitations and enabled latent space editing.

The human motion manifold can also be learned without labels, as shown by Holden et al. (2016), V-Poser (Pavlakos et al., 2019), and more recently the dedicated MoDi architecture (Raab et al., 2022). We show that our model is capable for such an unsupervised setting as well.

2.2 Diffusion Generative Models

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020) are a class of neural generative models, based on the stochastic diffusion process as it is modeled in Thermodynamics. In this setting, a sample from the data distribution is gradually noised by the diffusion process. Then, a neural model learns the *reverse process* of gradually denoising the sample. Sampling the learned data distribution is done by denoising a pure initial noise. Ho et al. (2020) and Song et al. (2020a) further developed the practices for image generation applications. For conditioned generation, Dhariwal & Nichol (2021), introduced classifier-guided diffusion, which was later on adapted by GLIDE (Nichol et al., 2021) to enable conditioning over CLIP textual representations. The Classifier-Free Guidance approach Ho & Salimans (2022) enables conditioning while trading-off fidelity and diversity, and achieves better results (Nichol et al., 2021). In this paper, we implement text-to-motion by conditioning on CLIP in a classifier-free manner, similarly to text-to-image (Ramesh et al., 2022; Saharia et al., 2022b). Local editing of images is typically defined as an inpainting problem, where a part of the image is constant, and the inpainted part is denoised by the model, possibly under some condition (Song et al., 2020b; Saharia et al., 2022a). We adapt this technique to edit motion's specific body parts or temporal intervals (in-betweening) according to an optional condition.

More recently, concurrent to this work, Zhang et al. (2022) and Kim et al. (2022) have suggested diffusion models for motion generation. Our work requires significantly fewer GPU resources and makes design choices that enable geometric losses, which improve results.

3 Motion Diffusion Model

An overview of our method is described in Figure 2. Our goal is to synthesize a human motion $x^{1:N}$ of length N given an arbitrary condition c. This condition can be any real-world signal that will dictate the synthesis, such as audio (Li et al., 2021; Aristidou et al., 2022), natural language (text-to-motion) (Tevet et al., 2022; Guo et al., 2022a) or a discrete class (action-to-motion) (Guo et al., 2020; Petrovich et al., 2021). In addition, unconditioned motion generation is also possible, which we denote as the null condition $c = \emptyset$. The generated motion $x^{1:N} = \{x^i\}_{i=1}^N$ is a sequences

Figure 2: **(Left) Motion Diffusion Model (MDM) overview.** The model is fed a motion sequence $x_t^{1:N}$ of length N in a noising step t, as well as t itself and a conditioning code c. c, a CLIP (Radford et al., 2021) based textual embedding in this case, is first randomly masked for classifier-free learning and then projected together with t into the input token z_{tk}. In each sampling step, the transformer-encoder predicts the final clean motion $\hat{x}_0^{1:N}$. **(Right) Sampling MDM.** Given a condition c, we sample random noise x_T at the dimensions of the desired motion, then iterate from T to 1. At each step t, MDM predicts the clean sample $\hat{x}_0$, and diffuses it back to x_{t-1}.

of human poses represented by either joint rotations or positions $x^i \in \mathbb{R}^{J \times D}$, where J is the number of joints and D is the dimension of the joint representation. MDM can accept motion represented by either locations, rotations, or both (see Section 4).

Framework. Diffusion is modeled as a Markov noising process, $\{x_t^{1:N}\}_{t=0}^T$, where $x_0^{1:N}$ is drawn from the data distribution and

$$q(x_t^{1:N}|x_{t-1}^{1:N}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}^{1:N}, (1-\alpha_t)I), \tag{1}$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. When α_t is small enough, we can approximate $x_T^{1:N} \sim \mathcal{N}(0, I)$. From here on we use x_t to denote the full sequence at noising step t.

In our context, conditioned motion synthesis models the distribution $p(x_0|c)$ as the reversed diffusion process of gradually cleaning x_T. Instead of predicting ϵ_t as formulated by Ho et al. (2020), we follow Ramesh et al. (2022) and predict the signal itself, i.e., $\hat{x}_0 = G(x_t, t, c)$ with the *simple* objective (Ho et al., 2020),

$$\mathcal{L}_{\text{simple}} = E_{x_0 \sim q(x_0|c), t \sim [1,T]}[\|x_0 - G(x_t, t, c)\|_2^2] \tag{2}$$

Geometric losses. In the motion domain, generative networks are standardly regularized using geometric losses Petrovich et al. (2021); Shi et al. (2020). These losses enforce physical properties and prevent artifacts, encouraging natural and coherent motion. In this work we experiment with three common geometric losses that regulate (1) positions (in case we predict rotations), (2) foot contact, and (3) velocities.

$$\mathcal{L}_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N} \|FK(x_0^i) - FK(\hat{x}_0^i)\|_2^2, \tag{3}$$

$$\mathcal{L}_{\text{foot}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)) \cdot f_i\|_2^2, \tag{4}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)\|_2^2 \tag{5}$$

In case we predict joint rotations, $FK(\cdot)$ denotes the forward kinematic function converting joint rotations into joint positions (otherwise, it denotes the identity function). $f_i \in \{0,1\}^J$ is the binary foot contact mask for each frame i. Relevant only to feet, it indicates whether they touch the ground, and are set according to binary ground truth data (Shi et al., 2020). In essence, it mitigates the foot-sliding effect by nullifying velocities when touching the ground.

Overall, our training loss is

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}. \tag{6}$$

Model. Our model is illustrated in Figure 2. We implement G with a straightforward transformer (Vaswani et al., 2017) encoder-only architecture. The transformer architecture is temporally aware, enabling learning arbitrary length motions, and is well-proven for the motion domain (Petrovich et al., 2021; Duan et al., 2021; Aksan et al., 2021). The noise time-step t and the condition code c are each projected to the transformer dimension by separate feed-forward networks, then summed to yield the token z_{tk}. Each frame of the noised input x_t is linearly projected into the transformer dimension and summed with a standard positional embedding. z_{tk} and the projected frames are then fed to the encoder. Excluding the first output token (corresponding to z_{tk}), the encoder result is projected back to the original motion dimensions, and serves as the prediction $\hat{x}_0$. We implement text-to-motion by encoding the text prompt to c with CLIP (Radford et al., 2021) text encoder, and action-to-motion with learned embeddings per class.

Sampling from $p(x_0|c)$ is done in an iterative manner, according to Ho et al. (2020). In every time step t we predict the clean sample $\hat{x}_0 = G(x_t, t, c)$ and noise it back to x_{t-1}. This is repeated from $t = T$ until x_0 is achieved (Figure 2 right). We train our model G using classifier-free guidance (Ho & Salimans, 2022). In practice, G learns both the conditioned and the unconditioned distributions by randomly setting $c = \emptyset$ for 10% of the samples, such that $G(x_t, t, \emptyset)$ approximates $p(x_0)$. Then, when sampling G we can trade-off diversity and fidelity by interpolating or even extrapolating the two variants using s:

$$G_s(x_t, t, c) = G(x_t, t, \emptyset) + s \cdot (G(x_t, t, c) - G(x_t, t, \emptyset)) \tag{7}$$

Editing. We enable motion in-betweening in the temporal domain, and body part editing in the spatial domain, by adapting diffusion inpainting to motion data. Editing is done only during sampling, without any training involved. Given a subset of the motion sequence inputs, when sampling the model (Figure 2 right), at each iteration we overwrite $\hat{x}_0$ with the input part of the motion. This encourages the generation to remain coherent to original input, while completing the missing parts. In the temporal setting, the prefix and suffix frames of the motion sequence are the input, and we solve a motion in-betweening problem (Harvey et al., 2020). Editing can be done either conditionally or unconditionally (by setting $c = \emptyset$). In the spatial setting, we show that body parts can be re-synthesized according to a condition c while keeping the rest intact, through the use of the same completion technique.

4 EXPERIMENTS

We implement MDM for three motion generation tasks: Text-to-Motion(4.1), Action-to-Motion(4.2) and unconditioned generation(5.2. Each sub-section reviews the data and metrics of the used benchmarks, provides implementation details, and presents qualitative and quantitative results. Then, we show implementations of motion in-betweening (both conditioned and unconditioned) and body-part editing by adapting diffusion inpainting to motion (5.1). Our models have been trained with $T = 1000$ noising steps and a cosine noise schedule. All of them have been trained on a single *NVIDIA GeForce RTX 2080 Ti* GPU for a period of about 3 days.

4.1 TEXT-TO-MOTION

Text-to-motion is the task of generating motion given an input text prompt. The output motion is expected to be both implementing the textual description, and a valid sample from the data distribution (i.e. adhering to general human abilities and the rules of physics). In addition, for each text prompt, we also expect a distribution of motions matching it, rather than just a single result. We evaluate our model using two leading benchmarks - KIT (Plappert et al., 2016) and HumanML3D (Guo et al., 2022a), over the set of metrics suggested by Guo et al. (2022a): *R-precision* and *Multimodal-Dist* measure the relevancy of the generated motions to the input prompts, *FID* measures the dissimilarity between the generated and ground truth distributions (in latent space), *Diversity* measures the variability in the resulting motion distribution, and *MultiModality* is the average variance given a single text prompt. For the full implementation of the metrics, please refer to Guo et al. (2022a). We use HumanML3D as a platform to compare different backbones of our model, discovering that the diffusion framework is relatively agnostic to this attribute. In addition, we conduct a user study comparing our model to current art and ground truth motions.

Figure 3: **Editing applications.** Light blue frames represent motion input and bronze frames are the generated motion. Motion in-betweening (left+center) can be performed conditioned on text or without condition by the same model. Specific body part editing using text is demonstrated on the right: the lower body joints are fixed to the input motion while the upper body is altered to fit the input text prompt.

Data. HumanML3D is a recent dataset, textually re-annotating motion capture from the AMASS (Mahmood et al., 2019) and HumanAct12 (Guo et al., 2020) collections. It contains $14,616$ motions annotated by $44,970$ textual descriptions. In addition, it suggests a redundant data representation including a concatenation of root velocity, joint positions, joint velocities, joint rotations and the foot contact binary labels. We also use in this section the same representation for the KIT dataset, brought by the same publishers. Although limited in the number $(3,911)$ and the diversity of samples, most of the text-to-motion research is based on KIT, hence we view it as important to evaluate using it as well.

Implementation. In addition to our Transformer encoder-only backbone (Section 3), we experiment MDM with three more backbones: (1) *Transformer decoder* injects z_{tk} through the cross-attention layer, instead of as an input token. (2) *Transformer decoder + input token*, where z_{tk} is injected both ways, and (3) *GRU* (Cho et al., 2014) concatenate z_{tk} to each input frame (Table 1). Our models were trained with batch size 64, 8 layers (except GRU that was optimal at 2), and latent dimension 512. To encode the text we use a frozen *CLIP-ViT-B/32* model. Each model was trained for $500K$ steps, afterwhich a checkpoint was chosen that minimizes the FID metric to be reported. Since foot contact and joint locations are explicitly represented in HumanML3D, we don't apply geometric losses in this section. We evaluate our models with guidance-scale $s = 2.5$ which provides a diversity-fidelity sweet spot (Figure 4).

Quantitative evaluation. We evaluate and compare our models to current art (JL2P Ahuja & Morency (2019), Text2Gesture (Bhattacharya et al., 2021), and T2M (Guo et al., 2022a)) with the metrics suggested by Guo et al. (2022a). As can be seen, MDM achieves state-of-the-art results in *FID*, *Diversity*, and *MultiModality*, indicating high diversity per input text prompt, and high-quality samples, as can also be seen qualitatively in Figure 1.

User study. We asked 31 users to choose between MDM and state-of-the-art works in a side-by-side view, with both samples generated from the same text prompt randomly sampled from the KIT test set. We repeated this process with 10 samples per model and 10 repetitions per sample. This user study enabled a comparison with the recent TEMOS model (Petrovich et al., 2022), which was not included in the HumanML3D benchmark. Fig. 4 shows that most of the time, MDM was preferred over the compared models, and even preferred over ground truth samples in 42.3% of the cases.

4.2 ACTION-TO-MOTION

Action-to-motion is the task of generating motion given an input action class, represented by a scalar. The output motion should faithfully animate the input action, and at the same time be natural and reflect the distribution of the dataset on which the model is trained. Two dataset are commonly used to evaluate action-to-motion models: HumanAct12 (Guo et al., 2020) and UESTC (Ji et al., 2018).

Method	R Precision (top 3)↑	FID↓	Multimodal Dist↓	Diversity→	Multimodality↑
Real	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
JL2P	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
Text2Gesture	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
T2M	$\mathbf{0.740^{\pm.003}}$	$1.067^{\pm.002}$	$\mathbf{3.340^{\pm.008}}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM (ours)	$0.611^{\pm.007}$	$\mathbf{0.544^{\pm.044}}$	$5.566^{\pm.027}$	$\mathbf{9.559^{\pm.086}}$	$2.799^{\pm.072}$
MDM (decoder)	$0.608^{\pm.005}$	$0.767^{\pm.085}$	$5.507^{\pm.020}$	$9.176^{\pm.070}$	$\mathbf{2.927^{\pm.125}}$
+ input token	$0.621^{\pm.005}$	$0.567^{\pm.051}$	$5.424^{\pm.022}$	$9.425^{\pm.060}$	$2.834^{\pm.095}$
MDM (GRU)	$0.645^{\pm.005}$	$4.569^{\pm.150}$	$5.325^{\pm.026}$	$7.688^{\pm.082}$	$1.2646^{\pm.024}$

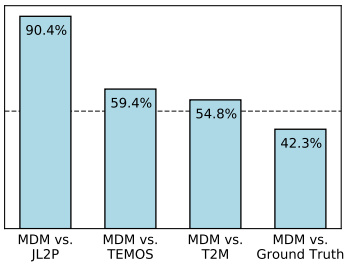
Table 1: **Quantitative results on the HumanML3D test set.** All methods use the real motion length from the ground truth. '→' means results are better if the metric is closer to the real distribution. We run all the evaluation 20 times (except *MultiModality* runs 5 times) and ± indicates the 95% confidence interval. **Bold** indicates best result.

Method	R Precision (top 3)↑	FID↓	Multimodal Dist↓	Diversity→	Multimodality↑
Real	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
JL2P	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Text2Gesture	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
T2M	$\mathbf{0.693^{\pm.007}}$	$2.770^{\pm.109}$	$\mathbf{3.401^{\pm.008}}$	$\mathbf{10.91^{\pm.119}}$	$1.482^{\pm.065}$
MDM (ours)	$0.396^{\pm.004}$	$\mathbf{0.497^{\pm.021}}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$\mathbf{1.907^{\pm.214}}$

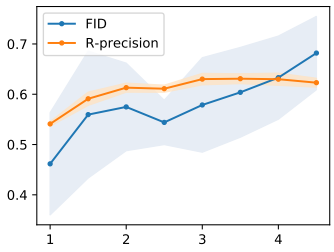
Table 2: **Quantitative results on the KIT test set.**

We evaluate our model using the set of metrics suggested by Guo et al. (2020), namely Fréchet Inception Distance (FID), action recognition accuracy, diversity and multimodality. The combination of these metrics makes a good measure of the realism and diversity of generated motions.

Data. HumanAct12 (Guo et al., 2020) offers approximately 1200 motion clips, organized into 12 action categories, with 47 to 218 samples per label. UESTC (Ji et al., 2018) consists of 40 action classes, 40 subjects and 25K samples, and is split to train and test. We adhere to the cross-subject testing protocol used by current works, with 225-345 samples per action class. For both datasets we use the sequences provided by Petrovich et al. (2021).



(a) KIT User Study



(b) Classifier-free scale sweep

Figure 4: **(a) Text-to-motion user study for the KIT dataset**. Each bar represents the preference rate of MDM over the compared model. MDM was preferred over the other models in most of the time, and 42.3% of the cases even over ground truth samples. The dashed line marks 50%. **(b) Guidance-scale sweep for HumanML3D dataset.** *FID* (lower is better) and *R-precision* (higher is better) metrics as a function of the scale s, draws an accuracy-fidelity sweet spot around $s = 2.5$.

Method	FID↓	Accuracy↑	Diversity→	Multimodality→
Real (INR)	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$
Real (ours)	$0.050^{\pm.000}$	$0.990^{\pm.000}$	$6.880^{\pm.020}$	$2.590^{\pm.010}$
Action2Motion (2020)	$0.338^{\pm.015}$	$0.917^{\pm.003}$	$6.879^{\pm.066}$	$2.511^{\pm.023}$
ACTOR (2021)	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$\mathbf{6.840^{\pm.030}}$	$2.530^{\pm.020}$
INR (2022)	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$
MDM (ours)	$0.100^{\pm.000}$	$\mathbf{0.990^{\pm.000}}$	$6.860^{\pm.050}$	$2.520^{\pm.010}$
w/o foot contact	$\mathbf{0.080^{\pm.000}}$	$\mathbf{0.990^{\pm.000}}$	$6.810^{\pm.010}$	$\mathbf{2.580^{\pm.010}}$

Table 3: **Evaluation of action-to-motion on the HumanAct12 dataset.** Our model leads the board in three out of four metrics. Ground-truth evaluation results are slightly different for each of the works, due to implementation differences, such as python package versions. It is important to assess the diversity and multimodality of each model using its own ground-truth results, as they are measured by their distance from GT. We show the GT metrics measured by our model and by the leading compared work, INR (Cervantes et al., 2022). **Bold** indicates best result, underline indicates second best, $\pm$ indicates 95% confidence interval, $\rightarrow$ indicates that closer to real is better.

Method	$\text{FID}_{\text{train}}$ ↓	FID_{test} ↓	Accuracy↑	Diversity→	Multimodality→
Real	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.001}$	$33.34^{\pm.320}$	$14.16^{\pm.06}$
ACTOR (2021)	$20.49^{\pm2.31}$	$23.43^{\pm2.20}$	$0.911^{\pm.003}$	$31.96^{\pm.33}$	$14.52^{\pm.09}$
INR (2022) (best variation)	$\mathbf{9.55^{\pm.06}}$	$15.00^{\pm.09}$	$0.941^{\pm.001}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$
MDM (ours)	$9.98^{\pm1.33}$	$\mathbf{12.81^{\pm1.46}}$	$0.950^{\pm.000}$	$33.02^{\pm.28}$	$\mathbf{14.26^{\pm.12}}$
w/o foot contact	$9.69^{\pm.81}$	$13.08^{\pm2.32}$	$\mathbf{0.960^{\pm.000}}$	$\mathbf{33.10^{\pm.29}}$	$\mathbf{14.06^{\pm.05}}$

Table 4: **Evaluation of action-to-motion on the UESTC dataset.** The performance improvement with our model shows a clear gap from state-of-the-art. **Bold** indicates best result, underline indicates second best, $\pm$ indicates 95% confidence interval, $\rightarrow$ indicates that closer to real is better.

Implementation. The implementation presented in Figure 2 holds for all the variations of our work. In the case of action-to-motion, the only change would be the substitution of the text embedding by an action embedding. Since action is represented by a scalar, its embedding is fairly simple; each input action class scalar is converted into a learned embedding of the transformer dimension.

The experiments have been run with batch size 64, a latent dimension of 512, and an encoder-transformer architecture. Training on HumanAct12 and UESTC has been carried out for $750K$ and $2M$ steps respectively. In our tables we display the evaluation of the checkpoint that minimizes the FID metric.

Quantitative evaluation. Tables 3 and 4 reflect MDM's performance on the HumanAct12 and UESTC datasets respectively. We conduct 20 evaluations, with 1000 samples in each, and report their average and a 95% confidence interval. We test two variations, with and without foot contact loss. Our model leads the board for both datasets. The variation with no foot contact loss attains slightly better results; nevertheless, as shown in our supplementary video, the contribution of foot contact loss to the quality of results is important, and without it we witness artifacts such as shakiness and unnatural gestures.

5 Additional Applications

5.1 Motion Editing

In this section we implement two motion editing applications - **in-betweening** and **body part editing**, both using the same approach in the temporal and spatial domains correspondingly. For **in-betweening**, we fix the first and last 25% of the motion, leaving the model to generate the remaining 50% in the middle. For **body part editing**, we fix the joints we don't want to edit and leave the

model to generate the rest. In particular, we experiment with editing the upper body joints only. In figure 3 we show that in both cases, using the method described in Section 3 generates smooth motions that adhere both to the fixed part of the motion and the condition (if one was given).

Method	FID↓	KID↓	Precision↑ Recall↑	Multimodality↑
ACTOR (2021)	48.80	0.53	0.72, 0.74	14.10
MoDi (2022)	**13.03**	**0.12**	**0.71**, **0.81**	**17.57**
MDM (ours)	31.92	0.36	0.66, 0.62	17.00

Table 5: **Evaluation of unconstrained synthesis on the HumanAct12 dataset.** We test MDM in the challenging unconstrained setting, and compare with MoDi (Raab et al., 2022), a work that was specially designed for such setting. We demonstrate that in addition to being able to support any condition, we can achieve plausible results in the unconstrained setting. **Bold** indicates best result.

5.2 Unconstrained Synthesis

The challenging task of unconstrained synthesis has been studied by only a few (Holden et al., 2016; Raab et al., 2022). In the presence of data labeling, e.g., action classes or text description, the labels work as a supervising factor, and facilitate a structured latent space for the training network. The lack of labeling make training more difficult. The human motion field possesses rich unlabeled datasets (Adobe Systems Inc., 2021), and the ability to train on top of them is an advantage. Daring to test MDM in the challenging unconstrained setting, we follow MoDi(Raab et al., 2022) for evaluation. We use the metrics they suggest (FID, KID, precision/recall and multimodality), and run on an unconstrained version of the HumanAct12 (Guo et al., 2020) dataset.

Data. Although annotated, we use HumanAct12 (see Section 4.2) in an unconstrained fashion, ignoring its labels. The choice of HumanAct12 rather than a dataset with no labels (e.g., Mixamo (Adobe Systems Inc., 2021)), is for compatibility with previous publications.

Implementation. Our model uses the same architecture for all forms of conditioning, as well as for the unconstrained setting. The only change to the structure shown in Figure 2, is the removal of the conditional input, such that z_{tk} is composed of the projection of t only. To simulate an unconstrained behavior, ACTOR Petrovich et al. (2021) has been trained by (Raab et al., 2022) with a labeling of one class to all motions.

Quantitative evaluation. The results of our evaluation are shown in table 5. We demonstrate superiority over works that were not designed for an unconstrained setting, and get closer to MoDi (Raab et al., 2022). MoDi is carefully molded for unconstrained settings, while our work can be applied to any (or no) constrain, and also provides editing capabilities.

6 Discussion

We have presented MDM, a method that lends itself to various human motion generation tasks. MDM is an untypical classifier-free diffusion model, featuring a transformer-encoder backbone, and predicting the signal, rather than the noise. This yields both a lightweight model, that is unburdening to train, and an accurate one, gaining much from the applicable geometric losses. Our experiments show superiority in conditioned generation, but also that this approach is not very sensitive to the choice of architecture.

A notable limitation of the diffusion approach is the long inference time, requiring about 1000 forward passes for a single result. Since our motion model is small anyway, using dimensions order of magnitude smaller than images, our inference time shifts from less than a second to only about a minute, which is an acceptable compromise. As diffusion models continue to evolve, beside better compute, in the future we would be interested in seeing how to incorporate better control into the generation process, and widen the options for applications even further.

ACKNOWLEDGEMENTS

We thank Rinon Gal for his useful suggestions and references. 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).

REFERENCES

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020.

Adobe Systems Inc. Mixamo, 2021. URL `https://www.mixamo.com`. Accessed: 2021-12-25.

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

Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pp. 565–574. IEEE, 2021.

A Aristidou, A Yiannakidis, K Aberman, D Cohen-Or, A Shamir, and Y Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 1–10. IEEE, 2021.

Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. *arXiv preprint arXiv:2203.13694*, 2022.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pp. 4346–4354, 2015.

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*, pp. 2021–2029, 2020.

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*, pp. 5152–5161, 2022a.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. *arXiv preprint arXiv:2207.01567*, 2022b.

Félix G Harvey and Christopher Pal. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia 2018 Technical Briefs*, pp. 1–4. 2018.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7134–7143, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

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.

Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *Proceedings of the 26th ACM international Conference on Multimedia*, pp. 1510–1518, 2018.

Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pp. 918–927. IEEE, 2020.

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

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

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pp. 5442–5451, October 2019.

Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2891–2900, 2017.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

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)*, pp. 10975–10985, 2019.

Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, pp. 10985–10995, October 2021.

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

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

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.

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*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022b.

Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.