Monkey See, Monkey Do: Harnessing Self-attention in Motion Diffusion for Zero-shot Motion Transfer

Anonymous Author(s)

Given the remarkable results of motion synthesis with diffusion models, a natural question arises: how can we effectively leverage these models for motion editing? Existing diffusion-based motion editing methods overlook the profound potential of the prior embedded within the weights of pre-trained models, which enables manipulating the latent feature space; hence, they primarily center on handling the motion space. In this work, we explore the attention mechanism of pre-trained motion diffusion models. We uncover the roles and interactions of attention elements in capturing and representing intricate human motion patterns, and carefully integrate these elements to transfer a leader motion to a follower one while maintaining the nuanced characteristics of the follower, resulting in zero-shot motion transfer. Manipulating features associated with selected motions allows us to confront a challenge observed in prior motion diffusion approaches, which use general directives (*e.g.*, text, music) for editing, ultimately failing to convey subtle nuances effectively. Our work is inspired by how a monkey closely imitates what it sees while maintaining its unique motion patterns; hence we call it ***Mo**nkey See, **Mo**nkey Do*, and dub it *MoMo*. Employing our technique enables accomplishing tasks such as synthesizing out-of-distribution motions, style transfer, and spatial editing. Furthermore, diffusion inversion is seldom employed for motions; as a result, editing efforts focus on generated motions, limiting the editability of real ones. MoMo harnesses motion inversion, extending its application to both real and generated motions. Experimental results show the advantage of our approach over the current art. In particular, unlike methods tailored for specific applications through training, our approach is applied at inference time, requiring no training. Our webpage, which includes links to videos and code, can be found at https://monkeyseedocg.github.io.

1 INTRODUCTION

Human motion synthesis is a fundamental task, useful for various fields including robotics, autonomous driving, health care, gaming and animation. Diffusion models [Ho et al. 2020; Sohl-Dickstein et al. 2015] stand out as the prevailing synthesis paradigm across different modalities, such as imaging [Saharia et al. 2022b], video [Ho et al. 2022], 3D point clouds [Luo and Hu 2021], and also motion [Dabral et al. 2023; Tevet et al. 2023]. With the emergence of foundation models [Bommasani et al. 2021], it has been natural for some modalities, like imaging and video, to evolve towards zero-shot editing [Geyer et al. 2024; Hertz et al. 2023]. Such works typically depend on the prior information encoded in the weights of pre-trained models, which facilitate latent feature space manipulation. This capitalizes on a deep understanding of their intricacies, with attention layers playing a dominant role in these methods.

However, the prior information encoded in pre-trained *motion* diffusion models remains largely unexplored. Furthermore, a notable disparity exists between the imaging and motion domains. Images possess a regularized 2D spatial structure with a considerable number of degrees of freedom (DoF), whereas our motion is defined over a 3D human skeleton with a 1D temporal axis, and significantly fewer DoF. Therefore, insights regarding pre-trained imaging models do not directly apply to motion.

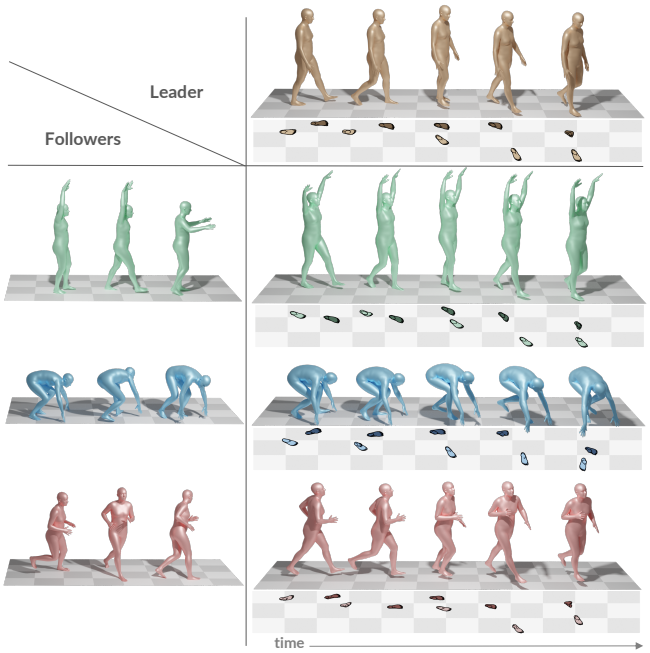

Fig. 1. **Motion transfer.** The top row displays a leader performing a walking motion. The left column showcases sample frames of four followers, each engaged in a different motion. The central block presents the output motion, where the outline of the leader (*e.g.*, leading leg) is transferred to the followers and integrated with their distinct motifs. Note the alignment of the steps for the leader and output motions. Our motion transfer is conducted by manipulating self-attention latent features in a zero-shot fashion.

In the following, the term *motion outline* denotes a structured plan that organizes the key blocks involved in a specific motion (*e.g.*, locomotion rhythm), and the term *motion motifs* denotes gestures or patterns (*e.g.*, characteristic pose). These terms are further elaborated in Sec. 3.

Human subjects exhibit highly expressive motions, containing a wealth of information. For example, a jumping motion can be combined with themes such as raising the arms or clapping the hands. A subtle change in each motion pattern would lead the viewer to a completely different impression. Similarly, discerning an individual's mood or age becomes possible by observing a short duration of their walking pattern. However, accurately conveying these subtle nuances through either high-level controls such as text or low-level controls such as joint trajectory poses a challenge. Moreover, motion datasets are limited, and acquiring real-world human motions representing various motifs solely through motion capture (MoCap) is expensive, slow, and unscalable.

Diffusion works that facilitate motion editing [Goel et al. 2023; Tevet et al. 2023] modify motion features (*e.g.*, rotation angles),

hence are limited to fixed sets of joints or operations. Furthermore, they rely on textual control and thus cannot convey subtle motifs. Works that *do* manipulate latent features [Raab et al. 2023; Tevet et al. 2022] precede the diffusion models era, lacking access to their rich priors.

In this work, we examine the *self-attention* mechanism within the motion domain (Sec. 4) and suggest an unpaired editing framework that transfers a leader motion to a follower one while preserving the subtle motifs of the follower, thereby reducing reliance on costly MoCap systems or overly general textual descriptions. We name our work "**Mo**nkey See, **Mo**nkey Do" (dubbed MoMo), as it encapsulates the concept of directly transferring a motion outline from a leader to a follower while retaining the follower's unique motifs, much like a monkey mimicking observed behavior in a monkeyish manner.

MoMo offers a versatile motion transfer technique, facilitating tasks unified by the core concept of transferring motifs from one motion to another. Special cases include style transfer (*e.g.*, transferring a walking motion to zombie walking), spatial editing (*e.g.*, transforming a jumping motion into "jumping with hands up"), action transfer (*e.g.*, transitioning from walking to running), and out-of-distribution synthesis (*e.g.*, generating a dancing gorilla). See Figs. 1 and 7 for more examples. Our model operates at inference time without requiring optimization or training and can function seamlessly with any underlying motion synthesis backbone (i.e., foundation model) that utilizes self-attention, regardless of its specific architecture.

Our study hypothesizes that self-attention elements can capture complex motion patterns, delves into the distinct functionalities of attention elements and examines their interplay. Inspired by research in the imaging domain [Alaluf et al. 2023; Cao et al. 2023; Hu et al. 2023], we devise a method where the capability of a query Q in a leader motion, is utilized to detect the most relevant keys K from a follower motion. Specifically, we calculate an attention score using the Q from the leader motion and the K from the follower motion. This score is utilized to extract a weighted combination of values V from the same follower motion. Consequently, a new motion is synthesized, incorporating the outline from the leader motion, and the motifs from the follower motion, maintaining fidelity to both. In essence, the leader motion determines the *"what"* and *"when"*, and the follower motion determines the *"how"*. For example, an instance of the "what" could be "a step forward with the right leg", while an instance of the "how" can be "running", "tiredly", or "with hands raised". Figure 5 illustrates the implicit semantic correspondence between the leader and follower motions, which does not require additional supervision.

Our work stands as the sole approach capable of utilizing motion DDIM inversion within diffusion models [Dhariwal and Nichol 2021; Song et al. 2020a]. In the field of imaging, the integration of inversion with diffusion is widely used, facilitating the manipulation of real images [Garibi et al. 2024; Huberman-Spiegelglas et al. 2023; Mokady et al. 2023]. However, in the motion domain, diffusion models typically avoid employing inversion. MoMo utilizes inversion, thereby enabling editing of both real and generated motions.

We introduce a comprehensive benchmark to evaluate our work. Our benchmark, named MTB, will be made publicly available. It comprises selected motion pairs from the HumanML3D [Guo et al. 2022] test set and is described in Sec. 5.1. Using MTB, we compare our model to current state-of-the-art methods and demonstrate that none offer the same breadth of functionality as MoMo, which consistently outperforms them.

2 RELATED WORK

2.1 Motion Synthesis

Multimodal synthesis. Petrovich et al. [2021, 2022] incorporate a Transformer [Vaswani et al. 2017] architecture for the tasks of action-to-motion and text-to-motion. T2M [Guo et al. 2022], T2M-GPT [Zhang et al. 2023b] and MotionGPT [Jiang et al. 2024] use VQ-VAE [Van Den Oord et al. 2017] to quantize motion, then sequentially synthesize it in the quantized space conditioned on text. More recently, MDM [Tevet et al. 2023] and Mofusion [Dabral et al. 2023] adapted the denoising diffusion framework [Ho et al. 2020] for motion synthesis and showed its merits for multimodal tasks such as action-to-motion, text-to-motion, and music-to-motion [Tseng et al. 2023]. The MAS [Kapon et al. 2023] algorithm extended diffusion to out-of-domain motions by leveraging video data. MoMo excels by using a *follower* reference instead of an overly generalized text.

Spatial control and editing. Motion Graphs [Arikan and Forsyth 2002; Kovar et al. 2002; Lee et al. 2002] are a popular data structure for traversing the pose space of a given motion dataset to synthesize new motion variations, often according to an input trajectory. GANimator [Li et al. 2022] and SinMDM [Raab et al. 2024] show that, analogously to images [Nikankin et al. 2023; Shaham et al. 2019], overfitting a single motion allows learning its internal distribution and synthesizing new variations of it with spatial and temporal variations. Using diffusion, MDM enables both joints and temporal editing by adapting image diffusion inpainting [Saharia et al. 2022a; Song et al. 2020b]. Followup works demonstrate the merits of propagating control signals through the gradual denoising process for various applications such as long motion synthesis [Petrovich et al. 2024; Zhang et al. 2023a], motion in-betweening [Cohan et al. 2024; Xie et al. 2023], and single joint control [Karunratanakul et al. 2023; Shafir et al. 2024]. ComMDM [Shafir et al. 2024] uses diffusion to synthesize two actors. InterGen [Liang et al. 2024] follows it and shows the effectiveness of cross-attention in the training process. SINC [Athanasiou et al. 2023] and FineMoGen [Zhang et al. 2024] use Large Language Models (LLMs) to break the text prompt to instruct each body part separately. Goel et al. [2023] use the coding skills of LLMs with predefined pose modifiers for frame editing, then blend them using diffusion. Our approach enables spatial editing of the *leader* motion as a special case of motion transfer.

Style transfer. One of the special cases enabled by MoMo is style transfer. Holden et al. [2017a, 2016] have suggested learning the motion manifold using an auto-encoder neural network. The latent space of the auto-encoder exposes semantic features of the motion, which enables motion stylization using the Gram matrix heuristic as suggested by Gatys et al. [2015]. Aberman et al. [2020] use the AdaIN heuristic to disentangle content and style as presented in StyleGAN [Karras et al. 2019], followed by Guo et al. [2024] and Kim et al. [2024]. Unlike MoMo, these models are trained on predefined styles and struggle to generalize.

2.2 Attention Control in the Imaging Domain

The latent information encapsulated in the attention layers of the popular UNet [Ronneberger et al. 2015] architecture is extensively used in the image domain to guide and control the denoising diffusion process. PnP [Tumanyan et al. 2023], MasaCtrl [Cao et al. 2023] and CIA [Alaluf et al. 2023] show that the self-attention layers encode structural information that can be used to edit an image without losing its original composition. Prompt-to-Prompt [Hertz et al. 2023] and Attend-and-Excite [Chefer et al. 2023] show that certain aspects of the image can be edited by manipulating the cross-attention with the input text, without affecting the rest. Patashnik et al. [2023] and Dahary et al. [2024] are manipulating the self- and cross-attention layers to control the layout of the image and avoid semantic leakage between its different parts. Tune-A-Video [Wu et al. 2023], TokenFlow [Geyer et al. 2024] and Q-NeRF [Patashnik et al. 2024] observe that the attention query, Q, encodes the structure while the key and value, K and V, encode the appearance, and use it for mutual editing of images preserving temporal and structural consistencies. Our work follows the latter, leveraging self-attention layers for motion editing. Unlike imaging works, our work uses layers from a transformer and not from a UNet.

3 MODEL

This section suggests an editing framework that transfers the outline of a leader motion to a follower one while preserving the motion motifs of the follower. Our unpaired framework operates at zero-shot, without requiring optimization or model training.

A motion *outline* denotes a structured plan that arranges the essential movements necessary for a specific motion. It provides a visual blueprint for comprehending the sequence of actions and transitions needed to execute the motion effectively. An example of a motion outline would be "stand still on frames 10-20, step with right leg on frames 21-25", etc. Note that the outline is as general as possible; for example, the type of step (walk, run, hop) belongs to the motifs. Motion *motifs* include subtle nuances, gestures, or patterns that convey meaning and emotion. These motifs may repeat and vary, forming expressive motion sequences, and aiding in establishing visual themes and narratives. Consider various running motions, each with distinct motifs. These motions convey personal styles expressed by body angle, foot positioning, hand gestures, airborne duration, etc. Even with extensive prompt engineering, capturing every subtle motif remains unattainable. Conversely, incorporating motifs from a given motion ensures complete fidelity to that motion.

Our framework enables transitions across motions of different temporal lengths. The outlines of a leader motion can be transferred to multiple followers, as shown in Fig. 1, and multiple outlines can be separately applied to a single follower.

3.1 Preliminaries

Motion Representation. Let N denote the number of frames in a motion sequence, and F denote the length of the features describing a single frame, also known as pose. Finally, let J denote the number of skeletal joints. We adhere to the representation used in the HumanML3D dataset [Guo et al. 2022], where the features from all the joints are concatenated into a single large feature, resulting in a motion representation $X \in \mathbb{R}^{N \times F}$. Details regarding the internal representation of the features can be found in the Appendix.

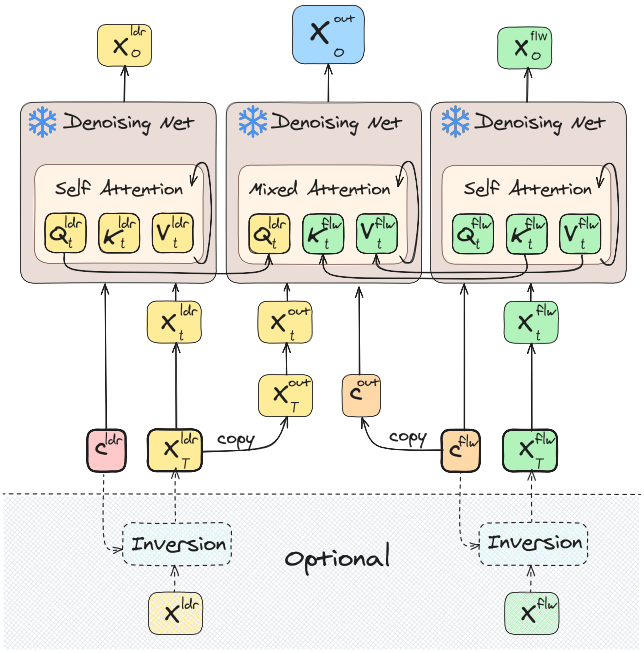

Fig. 2. **The MoMo Pipeline.** The input to our model is two noisy tensors, X_T^{ldr} and X_T^{flw}, produced by either inverting real motions or sampling a Gaussian noise. The two tensors represent leader and follower motions, and are given along with their associated text prompts. We initialize our output motion, X_T^{out}, using the initial noise from the leader motion and pair it with the text prompt from the follower motion. The three noised motions X_t^{ldr}, X_t^{flw} and X_t^{out}, are passed to the frozen denoising network at each timestep t, along with their prompts and with t. Within the denoising network, X_t^{out} undergoes mixed-attention by combining the query from the leader motion with the key and value from the follower motion. Meanwhile, X_t^{ldr} and X_t^{flw} follow a standard diffusion process.

Self Attention. We recap self-attention [Vaswani et al. 2017], as it plays a key role in our framework. Let IH ("input hidden") be a latent tensor of features fed as input to a self-attention layer, and let $\hat{H}$ be the output of this layer. The elements query, key, and value, are calculated respectively by

$$Q = IH \cdot W_q^T + b_q, \quad K = IH \cdot W_k^T + b_k, \quad V = IH \cdot W_v^T + b_v, \quad (1)$$

where (W_q, b_q), (W_k, b_k), and (W_v, b_v) are learned linear projections.

For each query vector $q_n \in Q$ at temporal location (*i.e.*, frame) n, an attention score is computed based on all keys in K. This score indicates the relevance of each key to the query q_n, assessing their similarity. The attention scores are normalized through a softmax operation, which determines the weighting of each value in V, to be used for updating the features at location n. The weighted values are then aggregated to produce the output at each query location,

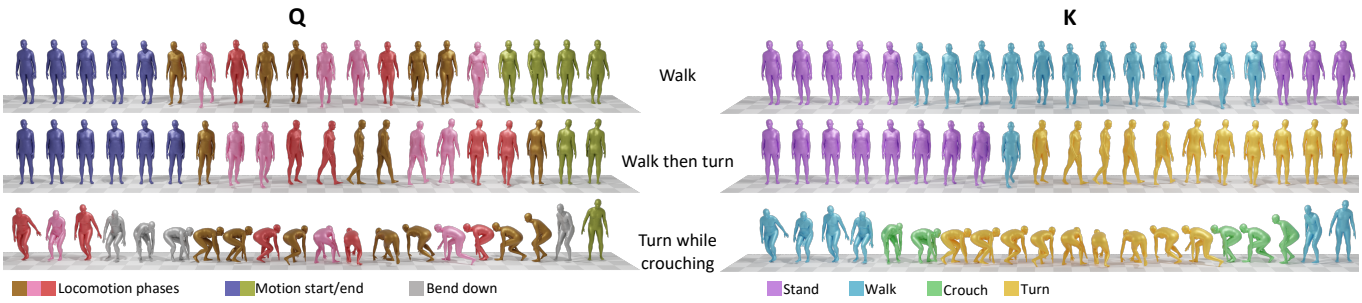

Fig. 3. **Dominant features in Q vs. K.** Each row depicts two copies of the same motion, showcasing the K-Means clustering of its Q and K features in the left and right columns, respectively. Note how the features in Q are dominated by the outline, while those in K are dominated by the motion motifs. In the Q column, periodic steps share clusters, ignoring unique patterns. In the K column, clusters are related to motion motifs; thus, walking, turning while walking, and crouching while walking have distinct clusters. Temporal information is evident in the clusters of Q but not in those of K. In the Q column, the beginnings of the first two motions and the end of all three are highlighted by the colors of low and high frame numbers, respectively.

via

$$A_n = \text{softmax}\left(\frac{q_n \cdot K^T}{\sqrt{|IH_n|}}\right), \qquad \hat{h}_n = A_n \cdot V, \tag{2}$$

where A_n is the normalized attention score at frame n, $\hat{h}_n \in \hat{H}$ is the self-attention result at frame n, and $|IH_n|$ is the features number in frame n. Finally, $\hat{H}$ is used as a residual and added to the input IH,

$$OH = IH + \hat{H}, \tag{3}$$

where OH ("output hidden") is the tensor passed to the next layer. We use multi-head attention and omit its notations for brevity.

MDM, DDPM and DDIM. Motion Diffusion Model (MDM) [Tevet et al. 2023] is a widespread model for human motion synthesis and editing. In our work we use a variation of it, hence we recap it here. MDM uses Denoising Diffusion Probabilistic Models (DDPM) [Ho et al. 2020], which are trained to transform unstructured noise into samples from a specified distribution. This is achieved through an iterative process involving the gradual removal of small amounts of Gaussian noise. Unlike MDM, in this work, we employ DDIM [Song et al. 2021] for inversion and deterministic inference, aiming to reconstruct inverted motions precisely to their original form. A recap regarding DDPM and DDIM can be found in the Appendix.

3.2 Pipeline

MoMo employs a pre-trained and fixed motion diffusion model to synthesize the output motion X^{out} by applying the motion outline of X^{ldr} onto X^{flw}, where X^{ldr} and X^{flw} are either given (real) motions or generated ones (See Fig. 2).

The input to our framework is two input noises, X_T^{ldr} and X_T^{flw} and their corresponding text prompts c^{ldr} and c^{flw}, respectively. The input noises are either inverted from real motions using DDIM [Song et al. 2021] or sampled from a Gaussian distribution, in which case MoMo runs on generated motions. The text prompts assist in controlling synthesis from noise and inverting the motions if inversion is used. At each timestep t, we pass the three noised motions X_t^{ldr}, X_t^{flw} and X_t^{out} to the denoising network. X_t^{ldr} and X_t^{flw} go through a standard DDIM denoising, while X_t^{out} is denoised using our mixed-attention block, described next. Finally, MoMo produces the output motion X_0^{out}.

The denoising network in our pipeline can be any motion-denoising model that utilizes self-attention layers. In our experiments, we employ a variant of MDM as a backbone. Note that, while related works in the imaging domain utilize the self-attention layers of a UNet [Ronneberger et al. 2015], we leverage the self-attention layers of a Transformer [Vaswani et al. 2017] due to the usage of MDM.

3.3 Leveraging Self-attention

Our proposed framework integrates self-attention components from both the leader and follower motions into a single output motion. In the imaging domain, Cao et al. [2023] have studied the self-attention layers of text-to-image denoising networks. They demonstrate that keeping the keys and values of these layers helps preserve the visual characteristics of objects when performing non-rigid manipulations on a given image. Alaluf et al. [2023] made further progress by combining structure and appearance from two images.

Inspired by these insights, this work illustrates the crucial functions of queries, keys, and values in encoding semantic motion information. We find that leveraging the queries, keys, and values from self-attention layers enables the transfer of semantic information across different motions.

In Sec. 4 we show that this approach enables the implicit transfer of motion patterns between semantically similar frames. More precisely, at each denoising step t, we use our mixed-attention block to inject the queries from the leader motion X^{ldr}, and the keys and values from the follower motion X^{flw} to the self-attention block of the output motion X^{out}, via

$$OH^{\text{out}} = IH^{\text{out}} + \text{softmax}\left(\frac{Q^{\text{ldr}} \cdot K^{\text{flw}^T}}{\sqrt{|IH_n|}}\right) V^{\text{flw}}. \tag{4}$$

4 UNDERSTANDING SELF-ATTENTION FEATURES

In this section, we explore some of the prior information encoded in pre-trained motion diffusion models and identify useful attributes within it. In the following, we demonstrate that (i) the *queries* Q establish a focal point for contextual determination, (ii) the *keys* K serve as a learned frame descriptor, enabling the model to assess the importance of different frames in the motion relative to a specific query, and (iii), the *values* V denote the contextual representation

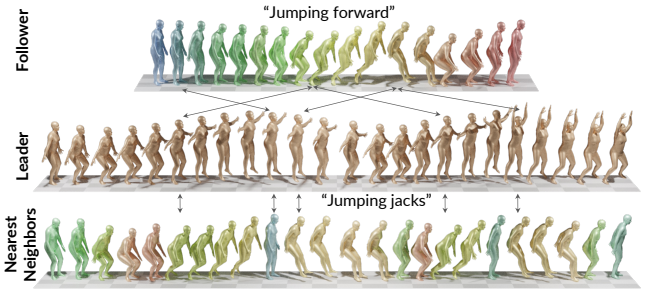

Fig. 4. **Correspondence via attention.** Follower frames are color-coded according to consecutive indices (top row). Nearest neighbor follower frames (bottom) are the ones that achieve the highest mixed-attention ($Q^{\text{ldr}} \cdot K^{\text{flw}^T}$) activation, shown respectively to leader's frames (middle row). These correspondences are semantically aligned, *e.g.*, moving "up" and "down" sub-motions are consistently assigned with follower moving "up" and "down" frames. Some of the nearest neighbors are highlighted with arrows.

we seek to generate, guiding the model in shaping the features at each query's temporal location. In particular, we show that while information regarding the outline and the motifs of the motion is contained in both Q and K, the outline information is more dominant in Q and the motifs information is more dominant in K. This key insight is presented in Sec. 4.1 and serves as the foundation on which our mixed-attention block was built; to the best of our knowledge, this insight has not been explored in the imaging or motion domains.

4.1 Distinguishing Between Q and K

To understand the different roles of queries and keys, we extract their features from a chosen self-attention layer ℓ at diffusion step t, reduce the dimension to d output channels by applying PCA, and group the frames into m clusters using K-Means. While Q and K contain both outline and motifs information, applying K-Means emphasizes the more dominant features. Figure 3 visualizes each cluster using a different color. Clustering K shows that it is dominated by the motion motifs and that similar motifs are grouped into the same clusters. Clustering Q depicts that it is dominated by the motion outline. In particular, (i) periodic sub-motions, like steps, are grouped into the same cluster, dominating over motion motifs, and (ii) the temporal signal is dominant. From (i) we conclude that the Qs in *different* periodic motions share similar features, thus they can be "understood" by Ks from other motions. This finding explains why our motion transfer works well. Interestingly, Q encodes the locomotion phases without explicitly learning it as in PFNNs [Holden et al. 2017b; Starke et al. 2022]. Our PCA is applied on 256 generated motions with 128 text prompts. We use ℓ=11, t=30, d=10, and m=10.

4.2 Correspondence via attention

Given two entirely different motions, we show that the nearest neighbor from a follower motion, for a query Q from a leader motion, would be the frame that resembles it most in its outline. In Fig. 4, for every frame n within the leader motion, we discern the frame in the follower motion that elicits the greatest activation in the attention map $q_n^{\text{ldr}} \cdot K^{\text{flw}^T}$. These pairings portray the transition from the leader to the follower motion, highlighting the most significant activations. As depicted, these pairings exemplify alignment in the outline, such as the synchronization of body movement ("up"/"down") across frames. The attention maps in this section are computed for layer 10 and diffusion step 50.

4.3 Attention Maps

After computing attention maps in the previous section, which revealed the highest activation for each frame, we now delve into the complete attention maps for two selected frames. Within each row in Fig. 5, we focus on a temporal region identified by the number of its center frame. In each column, we present the attention maps for each query location using various combinations of keys and queries. Naturally, when keys and queries from the same motion are multiplied ($Q^{\text{ldr}} \cdot K^{\text{ldr}^T}$), they produce high similarity scores for regions resembling the queried pose, especially the queried frame itself. In contrast, in the follower motion, the pose at the same frame number as the query may not be similar, making it unhelpful for retargeting the leader motion. In the right column, we use MoMo to compute $Q^{\text{ldr}} \cdot K^{\text{flw}^T}$. As illustrated, the queries focus on semantically similar regions in the follower motion.

Note that these correspondences are made despite the two motions having different sequences of poses. As a result, multiplying the attention maps with the follower values V^{flw}, enables the transfer of the outline of the leader motion onto the follower's motifs.

Finally, Fig. 6 depicts attention maps for the full motion length, providing visual insight into leader-follower correspondence.

5 EXPERIMENTS

The results in this work are computed using the transformer decoder version of MDM. The exact hyperparameter values are detailed in the Appendix. In practice, eq. (4) is applied in diffusion steps 90 to 10 (out of 100) and in layers 2 to 12 (out of 12).

5.1 Benchmark

To evaluate MoMo, we introduce ***M**otion **T**ransfer **B**enchmark*, dubbed *MTB*, of leader and follower motion pairs, which will be made available. MTB is a subset of the HumanML3D [Guo et al. 2022] test set and is straightforwardly filtered according to the textual prompts attached to the motions. For the leader motions, we include motions that contain locomotion verbs such as "run" or "walk". For the follower motions, we include motions with text that contains words suggesting motion motifs, such as "chicken" or "clap". The word choices, shared in our Appendix, are made straightforwardly and do not involve large language models.

We create pairs of leader and follower motions by combining sentences from the respective groups in a Cartesian product. However, such a combination induces approximately 46K pairs, which is more than we need. To decrease the number of pairs, instead of using all the cross combinations of leader and follower, we allow up to 20 repetitions of each follower sample and no repetitions for the leader samples. In practice, we use 4 leader search words, resulting in 842 motions, and 17 follower search words, resulting in 55 motions (some search terms result in less than three sentences). Altogether, MTB includes 842 (leader, follower) motion pairs.

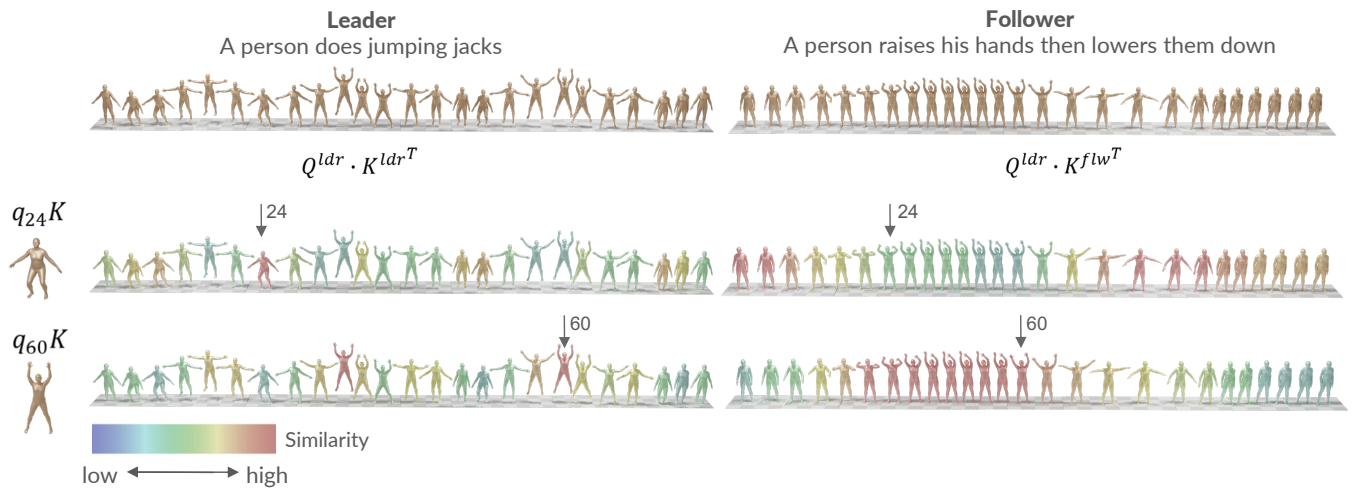

Fig. 5. **Attention map per query.** In the left column, we display three copies of the leader; in the right column, we show copies of the follower. The top copies depict the motions as they are, while the ones below highlight attention scores. We define two queries corresponding to different semantic temporal regions in the leader motion. Each query corresponds to a different pose, with varied arm direction or body stretch. Each motion column displays attention maps from a single layer, computed in different ways. In the left column, we present self-attention maps derived from queries and keys from the leader motion, causing each query to concentrate on semantically similar regions within that motion. The frame number related to each query is indicated with an arrow. For example, the query in frame 24 focuses on a pose of "standing low in an A pose", in the leader motion. However, frame number 24 corresponds to an entirely different pose in the follower motion in the right column. In the right column, we apply MoMo, aligning leader queries Q^{ldr} with follower keys K^{flw}. This way we ensure that each query from the leader motion aligns with semantically similar regions of the follower motion. For instance, in frame 60, the query highlights the region where the character raises their arms. The frames with higher correspondence (red) in the right column also belong to characters raising their arms.

5.2 Metrics

We aim to assess the results based on three criteria: (a) the quality of the output motions, (b) how closely the output motions follow the outline of the leader motion, and (c) how well the output motions align with the motifs of their associated follower motions. To evaluate these aspects, we have chosen specific metrics. Criterion (a) is evaluated using FID, (b) is assessed by foot contact similarity, and criterion (c) is evaluated by R-precision and similarity to follower locations and rotations. Details about these metrics are provided in the Appendix.

5.3 Baselines

No existing baseline encompasses all aspects of motion transfer addressed by MoMo. Hence, we compare MoMo with several works, each representing a different special case. We demonstrate that MoMo offers superior generalization capabilities compared to these works. Additionally, most prior art is tailored for specific applications through training, and often requires paired data. In contrast, MoMo is unpaired and accomplishes its functionalities during inference without necessitating specific training or optimization.

Motion transfer Using Naïve Nearest Neighbor. To illustrate the necessity of latent space editing, we compare MoMo with its naïve nearest neighbor (NN) equivalent, and examine three approaches. The first approach selects each frame's nearest counterpart in the follower motion and incorporates it into the output motion. The second computes a softmax on top of the similarity scores, extracting follower's frames according to the softmax weights. The third computes the similarity between the leader's Q and the follower's K latent features, and extracts the follower's value V according to the most similar (*i.e.*, nearest neighbor) frame.

Style Transfer. Style Transfer is a special case of motion transfer. It refers to doing a given action in a different way that represents an emotion or a physical state, such as "happily" or "like a monkey". Typically, it retains the original action (e.g., walk), while motion transfer encompasses any change in motion motifs, including changing the action (action transfer) or modifying a subset of joints without altering the style (spatial editing). While most motion style-transfer approaches [Aberman et al. 2020; Guo et al. 2024; Jang et al. 2022] excel with predefined style classes but struggle with unseen styles, our method seamlessly handles any given style motion. Existing style transfer methods require training and none of them utilizes diffusion models. We compare with state-of-the-art MoST [Kim et al. 2024], after we train it on the HumanML3D dataset.

Spatial Editing. Spatial editing adjusts specific joints, like the arms, while preserving the overall motion. Among spatial editing motion diffusion methods, MDM inpainting [Tevet et al. 2023], as well as MEO [Goel et al. 2023], use textual control and modify end features of a motion (e.g., rotation angles) without utilizing latent space. MDM inpainting excels at editing broad sets of joints, but needs refinement for individual joint edits [Shafir et al. 2024]. It relies on hard-coded joint masks, limiting edits to predefined body parts. MEO [Goel et al. 2023] uses a finite hard-coded set of editing operations. In contrast, our approach is not limited to editing a large set of joints, hard-coded masks, or a finite set of editing operations. MDM Inpainting is the closest diffusion spatial editing work that provides code. It expects an input motion and a text, so for each pair

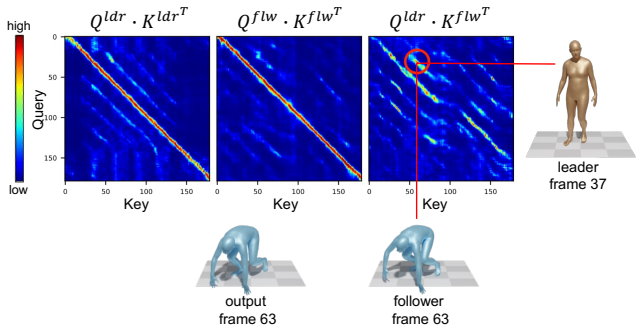

Fig. 6. **Attention maps for full motions.** The left and middle self-attention maps relate to a leader and a follower, performing "walking" and "walking like a gorilla" motions, respectively. To the right is their mixed-attention map. For self-attention, the main diagonal indicates self-correlation and the secondary diagonals indicate the periodicity in the walks. For mixed-attention, the high attention diagonals relate to the correlation between the leader and the follower. In the frames attributed to the circled high attention score within the mixed-attention map, both leader and follower depict a similar stepping pose, justifying the high activation. The resulting output frame is similar, but not identical, to the follower's, attributed to the weighted sum that combines multiple inputs. All maps relate to layer 6 and diffusion step 70 in our backbone.

of benchmark motions, the leader is used as is, and the follower is given by its text prompt.

OOD Synthesis. Out-of-distribution (OOD) synthesis entails generating motions that were not encountered during the training phase, posing a challenge to the network's generalization capabilities. For instance, when attempting to generate a dancing gorilla, the network struggles to generalize to this combination despite being trained on individual instances of dancing humans and of non-dancing gorillas. This difficulty arises from the sparse representation of such combinations in the latent space. By applying our proposed technique, we generate latent features located sparsely away from all other features, and yet, they are denoised into a natural motion. Existing motion diffusion methods do not transfer a given motion into an OOD one. Hence, we evaluate the OOD samples in our benchmark using style transfer and spatial editing baselines.

Action Transfer. Action transfer is the application of transferring the outline of one action into another, say, "running to walking" or "walking to crawling", where the output would be doing the follower's action in the same rhythm and leg order as the leader. As for OOD synthesis, no current method diffuses motion into a different action. We assess action transfer samples in our benchmark using style transfer and spatial editing baselines.

5.4 Quantitative Results

In Tab. 1 we compare MoMo with the baselines mentioned above, using the MTB benchmark. Our work excels in FID and R-precision while achieving comparable similarity scores. However, baselines with the highest similarity scores show inferior FID and R-precision results. MoMo's version operating on generated motions outperforms the one on inverted motions.

5.5 Qualitative Results

Our supplementary video reflects the quality of our results. It presents multiple transferred motions and comparisons to other baselines. It shows that the first nearest neighbor approach results in jittery outcomes, the second produces unnatural outcomes with foot sliding, MDM Inpainting cannot generalize beyond its spatial editing expertise, and MoST struggles to generalize on unseen styles. MoMo offers a versatile motion transfer technique expressed in Figs. 1 and 7 where we see how it facilitates tasks unified by the core concept of transferring a motion outline from a leader to a follower while retaining the follower's unique motifs. Arranged from top to bottom, Fig. 1 demonstrates special cases of spatial editing, style transfer, and action transfer. Fig. 7 illustrates action transfer, out-of-distribution synthesis, and another action transfer.

Table 1. **Comparison with baselines on the MTB benchmark**. MoMo exhibits the best FID and R-precision results and comparable similarity scores. MoMo's version that runs on generated motions (abbreviated "Gen.") has slightly better results than the one that runs on inverted ones ("Inv."). Baselines that exhibit the highest similarity metrics scores, show poor FID and R-precision results. Naïve nearest neighbor variations attain good similarity to the follower, as they are built by features copied from it. However, they exhibit poor FID due to a jittery output. MoST results are too similar to the leader, hence fail in similarity to the follower, and MDM Inpainting works well only when the upper body part needs to be edited, due to its fixed mask. **Bold** and underline indicate best and second-best results, respectively.

Model \ Metric	FID ↓	R Precision (top 3) ↑	Leader Foot Contact Sim. ↑	Follower Rot. Sim. ↑	Follower Loc. Sim. ↑
NN motion space	6.17	0.385	0.830	**1.00**	**1.00**
+ softmax	11.9	0.312	0.756	0.994	0.986
NN latent space	3.63	0.384	0.798	0.981	0.966
MoST [2024]	15.2	0.240	0.824	0.207	0.227
MDM Inp. [2023]	3.51	0.213	**1.00**	0.244	0.329
MoMo Gen. (Ours)	**2.33**	0.439	0.816	0.993	0.972
MoMo Inv. (Ours)	2.50	**0.490**	0.793	0.933	0.856

Table 2. **Layers and steps ablation.** The table displays representative results of the variations we experimented with diffusion steps and layers in which the self-attention features are manipulated. The top row displays the configuration of our selected model, where we apply mixed-attention for self-attention layers 2-12 and for diffusion steps 90 to 10 (out of 100). In the middle block, we maintain our best layer configuration and test various diffusion step ranges. In the bottom block, we maintain our best diffusion steps configuration and experiment with the range of layers. To select the best configuration, we prioritized FID and R-Precision over the other metrics.

Layers	Diffusion Steps	FID ↓	R Precision (top 3) ↑	Leader Foot Contact Sim. ↑	Follower Rot. Sim. ↑	Follower Loc. Sim. ↑
2 - 12	10 - 90	**2.334**	**0.439**	0.816	**0.993**	0.972
2 - 12	20 - 80	3.028	0.412	0.820	**0.993**	0.973
2 - 12	15 - 90	2.833	0.406	0.817	0.992	**0.975**
2 - 12	20 - 70	2.867	0.416	0.821	0.991	0.973
4 - 9	10 - 90	4.063	0.373	0.795	0.978	0.971
5 - 11	10 - 90	2.971	0.393	**0.837**	0.989	0.967
4 - 10	10 - 90	3.098	0.404	0.821	0.991	0.974

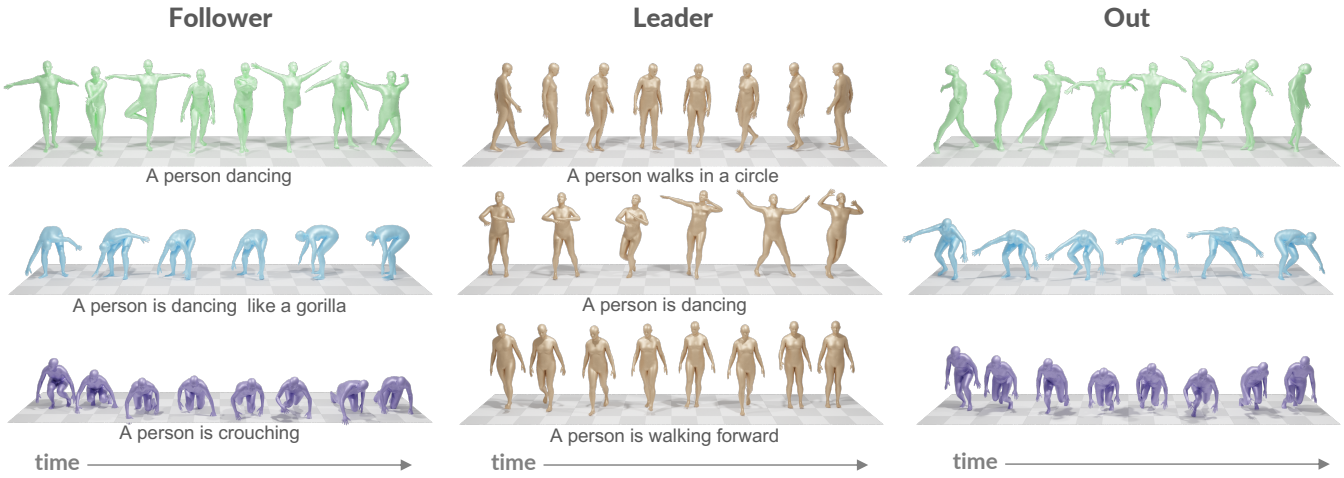

Fig. 7. **Qualitative Results.** The results here demonstrate the transfer of the leader motion's outline, such as the sequence and pace of steps, to the follower while preserving the subtle nuances of the follower's motion. The first and third rows relate to the special case of an action transfer. In the second row, we illustrate an out-of-distribution synthesis scenario; initially, the follower lacks a dancing motion, despite the text prompt. However, after the transfer process, the synthesized motion imitates a gorilla dancing, as specified.

5.6 Ablation

Our initial study focuses on identifying the appropriate layers and diffusion steps for which mixed-attention should be applied. We denote the range of layers and diffusion steps where mixed-attention is applied as [*s-layer*, *e-layer*] and [*s-step*, *e-step*], respectively. We have tested approximately 200 configurations, with varying values of *s-layer*, *e-layer*, *s-step*, and *e-step*. Table 2 displays representative results of the variations we experimented with.

Another ablation test examines the textual prompt used during the synthesis of the output motion. Results in Tab. 3 confirm that our choice to replicate the follower's prompt is optimal.

6 CONCLUSION, LIMITATIONS AND FUTURE WORK

We have explored and leveraged the powerful mechanism of self-attention, within pre-trained motion diffusion models. Our study has resulted in a novel approach for motion editing in diffusion models, which we have demonstrated through motion transfer.

Our zero-shot and unpaired methodology, named MoMo (**Mo**nkey See, **Mo**nkey Do), demonstrates promising capabilities in transferring the outline of a leader motion onto a follower one while preserving the subtle motifs of the follower. This enables a variety of sub-tasks such as out-of-distribution motion synthesis, style transfer, action transfer, and spatial editing. By harnessing motion inversion, MoMo extends its applicability to both real and generated motions.

Experimental results demonstrate the merits of our approach compared to the baselines. Notably, it excels in applying functionalities at inference time without additional training. Our findings shed light on the importance of attention in human motion and lay the groundwork for future advancements in the field.

The primary limitation of our work is the difficulty in transferring a motion when basic outline elements of the leader motion are lacking in the follower. For example, if the leader motion depicts walking while the follower remains stationary, then the output motion will also be stationary, receiving no input from the leader. However, this limitation may be inherent to the task description itself, as our objective is primarily to transfer motions that share some degree of commonality in their outlines. Another current limitation is that once a leader and a follower motions are determined, there is no diversity in the output motions obtained from them. This limitation arises from the deterministic nature of the DDIM [Song et al. 2020a] diffusion model we are using. In the future, output diversity could be achieved using non-deterministic models (e.g., DDPM).

A possible future direction we offer here is to explore latent feature layers other than self-attention, *e.g.*, cross-attention and feed-forward. Additionally, our findings on motion motifs preservation could be applied to personalization applications, a domain popular in imaging [Gal et al. 2023] but currently lacking in the motion domain.

Table 3. **Text prompt ablation.** In each row, a different textual prompt is utilized for the output motion. Results indicate that using the same text as the follower yields the best outcomes.

Prompt \ Metric	FID ↓	R Precision (top 3) ↑	Leader Foot Contact Sim. ↑	Follower Rot. Sim. ↑	Follower Loc. Sim. ↑
Same as follower	**2.572**	**0.434**	0.814	**0.994**	**0.975**
None	3.182	0.410	0.817	0.986	0.948
"A person"	3.282	0.391	**0.824**	0.986	0.947

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64–1.

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Cross-Image Attention for Zero-Shot Appearance Transfer. arXiv:2311.03335 [cs.CV]

Okan Arikan and David A Forsyth. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 483–490.

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2023. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 9984–9995.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. 2024. Flexible Motion In-betweening with Diffusion Models.

Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2024. Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation.

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

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=NAQvF08TcyG

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. ReNoise: Real Image Inversion Through Iterative Noising.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A Neural Algorithm of Artistic Style. arXiv:1508.06576 [cs.CV]

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2024. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=daEqXJ0yZo

Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. 2023. Iterative Motion Editing with Natural Language.

Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. 2024. Generative Human Motion Stylization in Latent Space. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=daEqXJ0yZo

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 5152–5161.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross attention control. In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, OpenReview.net.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

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

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

Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. 2017a. Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37, 4 (2017), 42–49.

Daniel Holden, Taku Komura, and Jun Saito. 2017b. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

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

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2023. An Edit Friendly DDPM Noise Space: Inversion and Manipulations.

Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. 2022. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)* 41, 3 (2022), 1–16.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2024. MotionGPT: Human Motion as a Foreign Language. *Advances in Neural Information Processing Systems* 36 (2024).

Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H. Bermano. 2023. MAS: Multi-view Ancestral Sampling for 3D motion generation using 2D diffusion. arXiv:2310.14729 [cs.CV]

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 4401–4410.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 2151–2162.

Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. 2024. MoST: Motion Style Transformer between Diverse Action Contents.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion graphs. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques (SIGGRAPH 2002)*. ACM, New York, NY, USA, 473––482.

Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard. 2002. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. ACM New York, NY, USA, New York, NY, USA, 491–500.

Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. 2022. GANimator: Neural Motion Synthesis from a Single Sequence. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 138.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. In *International Journal of Computer Vision*. Springer, Berlin/Heidelberg, Germany, 1–21.

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

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, 6038–6047.

Yaniv Nikankin, Niv Haim, and Michal Irani. 2023. SinFusion: Training Diffusion Models on a Single Image or Video. In *International Conference on Machine Learning*. PMLR, PMLR, PMLR, 26199–26214.

Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De la Torre. 2024. Consolidating Attention Features for Multi-view Image Editing.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing Object-Level Shape Variations with Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Springer International Publishing, Berlin/Heidelberg, Germany, 23051–23061.

Mathis Petrovich, Michael J. Black, and Gül Varol. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Washington, DC, USA, 10985–10995.

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

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. 2024. STMC: Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation.

Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. 2023. MoDi: Unconditional Motion Synthesis from Diverse Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 13873–13883.

Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. 2024. Single Motion Diffusion. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, OpenReview.net.

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

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022b. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494. https://openreview.net/forum?id=08Yk-n5l2Al

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. 2024. Human Motion Diffusion as a Generative Prior. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. SinGAN: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 4570–4580.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=St1giarCHLP

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

Sebastian Starke, Ian Mason, and Taku Komura. 2022. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. 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*. Springer, Springer International Publishing, Berlin/Heidelberg, Germany, 358–374.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, OpenReview.net. https://openreview.net/forum?id=SJ1kSyO2jwu

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 448–458.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 1921–1930.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

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

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Washington, DC, USA, 7623–7633.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2023. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, OpenReview.net.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023b. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. 2024. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems* 36 (2024).

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming yu Liu. 2023a. DiffCollage: Parallel Generation of Large Content with Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, IEEE Computer Society, Washington, DC, USA, 10188–10198.

APPENDIX

This Appendix adds details on top of the ones given in the main paper. While the main paper stands on its own, the details given here may shed more light.

In appendix A we provide more details regarding the preliminaries of our work: motion representation, and the models MDM, DDPM, and DDIM. Appendix B elaborates on our experiments, in particular the benchmark, metrics, hypermparameters, and additional implementation details.

A PRELIMINARIES - MORE DETAILS

Motion Representation. Recall that N denotes the number of frames in a motion sequence, F denotes the length of the features describing a single frame, J denotes the number of skeletal joints, and $X \in \mathbb{R}^{N\times F}$ denotes a motion. Each feature is redundantly represented with the joint angles, positions, velocities, and foot contact [Guo et al. 2022]. Each single pose is defined by

$$(\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^p, j^r, j^v, c^f) \in \mathbb{R}^F,$$

where $\dot{r}^a \in \mathbb{R}$ is the root angular velocity along the Y-axis. $\dot{r}^x, \dot{r}^z \in \mathbb{R}$ are root linear velocities on the XZ-plane, and $r^y \in \mathbb{R}$ is the root height. $j^p \in \mathbb{R}^{3(J-1)}$, $j^r \in \mathbb{R}^{6(J-1)}$ and $j^v \in \mathbb{R}^{3J}$ are the local joint positions, velocities, and rotations relative to the root, and $c^f \in \mathbb{R}^4$ are binary features denoting the foot contact labels for four foot joints (two for each leg).

MDM, DDPM and DDIM. Recall that MDM [Tevet et al. 2023] does human motion synthesis and editing and that it uses DDPMs [Ho et al. 2020]. In the following, we briefly describe the mechanism of DDPM.

An input motion x_0, is subjected to a Markov noise process consisting of T steps, resulting in the sequence $\{x_t\}_{t=0}^T$, such that

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I), \tag{5}$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. When α_t is small enough, we can approximate $x_T \sim \mathcal{N}(0, I)$.

x_0 can be modeled via the reversed diffusion process by gradually cleaning x_T, using a generative network p_θ. MDM [Tevet et al. 2023] predicts the input motion, denoted $\hat{x}_0$, rather than ϵ_t, such that $\hat{x}_0 = p_\theta(x_t, t)$. Then, the widespread diffusion loss is applied:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t\sim[1,T]}\|x_0 - p_\theta(x_t, t)\|_2^2. \tag{6}$$

During inference, synthesis iterates from pure noise x_T. In each iteration, the denoising network p_θ predicts a clean version of the current sample x_t. The predicted clean sample $\hat{x}_0$ is then "re-noised" to create the next sample x_{t-1}, repeatedly until $t = 0$.

Denoising Diffusion Implicit Models (DDIM) [Song et al. 2021] enable a non Markovian, deterministic, version of DDPMs. DDIM can be applied on a network pre-trained with DDPM, for either accelerated inference, or inversion. In this work, we employ DDIM for inversion and deterministic inference, to reconstruct inverted motions precisely to their original form.

B EXPERIMENTS - MORE DETAILS

B.1 Benchmark

The complete list of filter terms used to create the MTB benchmark is as follows. For each pair in the benchmark, the leader contains one of the following locomotion key words: "run", "walk", "jump", "danc". The follower contains one of the aforementioned locomotion key words, plus one style or action key word, from the list "gorilla", "drunk", "robot", "chicken", "frog", "monkey", "style", "like", "old", "child", "raise", "clap", "wav", "kick", "punch", "push", "pull".

Following are several examples of pairs from the MTB benchmark. Each example is identified by its text prompt and its index number in the HumanML3D dataset.

- Leader: "a person walks towards the left making a wide 's' shape" (#009488). Follower: "a person walks in a clockwise circle and raises their hand to their face to yawn" (#004222).
- Leader: "a person runs and then jumps" (#007291). Follower: "the drunk guy struggles to walk down the street" (#005037).
- Leader: "it is a person walking backwards" (#007199). Follower: "person is walking with his arms out like he is balancing" (#010823).

B.2 Metrics

FID [Guo et al. 2022; Heusel et al. 2017]. We use this metric to assess the quality of the output motions. The Fréchet Inception Distance (FID) evaluates the similarity between the distribution of real motions and synthesized ones We consider the benchmark dataset as the ground truth distribution and extract features from both real motions in the benchmark and generated output motions. Motions are deemed of high quality if they exist within the manifold of the ground truth.

R-precision [Guo et al. 2022]. We use this metric to assess the alignment between each output motion and the motion motifs of their associated follower motions. R-precision uses a latent space shared between text and motion to measure the distance between motion and text embeddings. Hence, it measures whether a motion reflects some text prompt. The text prompt of each output motion is copied from the follower used when generating it, so a successful match means the output motion adheres to the motion pattern of the follower.

For each generated motion, we create a description pool consisting of its ground-truth text description and 31 randomly chosen, unrelated descriptions from the test set. We then calculate and rank the Euclidean distances between the motion feature and the text feature of each description in the pool. We measure success by the average accuracy at the top-1, top-2, and top-3 positions. A successful retrieval occurs when the ground truth entry ranks in the top-k candidates.

Foot Contact Similarity. This metric assesses how well the output motion retains the locomotion rhythm of the leader motion. Successful motion transfer should preserve the rhythm from leader to output motion, so the foot contact labels of the two motions should closely align. We measure this similarity score by comparing the foot contact data within the HumanML3D motion features. Matching foot contact labels are counted as hits while differing labels are counted as misses. The metric score is determined by the rate of hits.

Similarity to Follower Rotations. This metric evaluates the fidelity of the output motion to the subtle motifs found in the follower. Specifically, it assesses the resemblance of rotations within the output motion to either the leader or follower motions. Given that subtle nuances are typically expressed in most joints' rotations, we anticipate observing a higher degree of similarity to the follower motion.

To compute this metric, the initial step is to identify, for each frame, the nearest neighbor in both the leader and follower motions. This entails identifying the closest match for each frame.

Once the nearest neighbors are identified, the metric calculates the rate of frames where the similarity to the follower's nearest neighbor surpasses that of the leader's. Essentially, it evaluates the proportion of rotation frames where the outbound motion exhibits greater alignment with the follower's motion compared to the tracked one.

Similarity to Follower Locations. This metric also evaluates the fidelity of the output motion to follower's motifs. Similar to the previous metric, it calculates the nearest neighbors for each frame, but this time for the locations relative to the root. As with the previous metric, we expect a greater degree of similarity to the follower motion.

B.3 Implementation Details

Table 4. **Hyperparameters used for our backbone.** Our backbone is a variation of MDM, with the hyperparameters listed here.

Name	Value
Model	
Architecture	Transformer Decoder
layers	12
latent dim	512
Diffusion	
diffusion steps	100
noise schedule	cosine
guidance scale	2.5
Training	
batch size	32
lr	0.0001
dropout	0.1
num steps	600000
warmup steps	0
weight decay	0
seed	10

Table 5. **Hyperparameters used during inference when applying MoMo.**

Name	Value
Applying mixed-attention	
Layers	range(1, 12)
Diffusion steps	range(10, 91)
Text prompt	
Outputmotion	Same as follower

Backbone and hyperparameters. The backbone we use for reporting experimental results is the transformer decoder variation of MDM [Tevet et al. 2023]. We retrain it on the HumanML3D [Guo et al. 2022] dataset, with the hyperparameters shown in Tab. 4. In particular, we slightly change the architecture such that the text embedding is given twice: once as an extra-temporal token like

in the transformer encoder variation, and once as separate word tokens using cross-attention.

Additionally, Tab. 5 provides hyperparameters used during inference when applying MoMo.

Controlling directions during motion transfer. A natural challenge arises when the locomotion direction of the leader motion differs from that of the follower's. In such cases, the output motion would retain the outline of the leader, such as the timing of steps, but would proceed in a different direction. Fortunately, the solution is straightforward. For generated motions, we create multiple followers, each generated with a different seed, for the given text prompt. MoMo is then applied to the concatenation of all followers together. Our experiments show that generating just a few followers ensures that the output motion closely follows the direction of the follower. For inverted motions, using different seeds would not help as the inversion by DDIM is deterministic. Our solution is to rotate the given motion around the vertical axis in several rotation angles. Similar to the generative case, we apply our framework to the concatenation of all rotated followers. Consequently, each leader's frame attends to the frames of all the followers and achieves the highest attention scores with those that have directions similar to its own. This facilitates near-accurate direction transfer. In practice, we have found that this solution is nearly equivalent to copying the root rotation value from the leader to the output motion. For evaluation, we used the latter, to accelerate computation time.