MotionMix: Weakly-Supervised Diffusion for Controllable Motion Generation

Nhat M. Hoang[1,2*], **Kehong Gong**[1†], **Chuan Guo**[1*], **Michael Bi Mi**[1]

[1]Huawei Technologies Co., Ltd.,
[2]Nanyang Technological University
nhat005@e.ntu.edu.sg, gongkehong@u.nus.edu, cguo2@ualberta.ca, michaelbimi@yahoo.com

Abstract

Controllable generation of 3D human motions becomes an important topic as the world embraces digital transformation. Existing works, though making promising progress with the advent of diffusion models, heavily rely on meticulously captured and annotated (e.g., text) high-quality motion corpus, a resource-intensive endeavor in the real world. This motivates our proposed **MotionMix**, a simple yet effective weakly-supervised diffusion model that leverages both *noisy* and *unannotated* motion sequences. Specifically, we separate the denoising objectives of a diffusion model into two stages: obtaining conditional rough motion approximations in the initial $T - T^*$ steps by learning the noisy annotated motions, followed by the unconditional refinement of these preliminary motions during the last T^* steps using unannotated motions. Notably, though learning from two sources of imperfect data, our model does not compromise motion generation quality compared to fully supervised approaches that access gold data. Extensive experiments on several benchmarks demonstrate that our MotionMix, as a versatile framework, consistently achieves state-of-the-art performances on text-to-motion, action-to-motion, and music-to-dance tasks.

1 Introduction

The rapidly arising attention and interest in digital humans bring up the great demand for human motion generation, in a wide range of fields such as industrial game and movie animation (Ling et al. 2020), human-machine interaction (Koppula and Saxena 2013), VR/AR and metaverse development (Lee et al. 2021). Over the years, automated generation of human motions that align with user preferences, spanning aspects such as prefix poses (Ruiz, Gall, and Moreno-Noguer 2018; Guo et al. 2022c), action classes (Petrovich, Black, and Varol 2021; Cervantes et al. 2022), textual descriptions (Petrovich, Black, and Varol 2022; Ahuja and Morency 2019; Tevet et al. 2022), or music (Aristidou et al. 2021; Siyao et al. 2022; Gong et al. 2023), has been a focal point of research. Recently, building upon the advancement of diffusion models, human motion generation has experienced a notable improvement in quality and controllability.

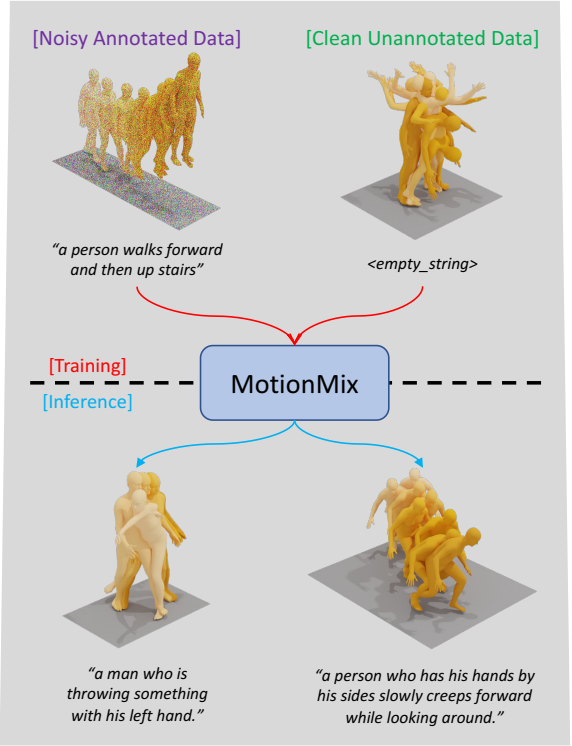

Figure 1: Example of applying MotionMix on text-to-motion task. Unlike prior works, our training data are comprised of *noisy annotated motions* and *unannotated motions*.

However, these prior diffusion models are commonly trained on well-crafted motions that come with explicit annotations like textual descriptions. While capturing motions from the real world is a laborious effort, annotating these motion sequences further urges the matter.

In contrast, motions with lower fidelity or fewer annotations are more accessible in the real world. For example, 3D human motions are readily extracted from monocular videos through video-based pose estimation (Kanazawa et al. 2017; Kocabas, Athanasiou, and Black 2019; Choutas et al. 2020). Meanwhile, a wealth of unannotated motion sequences, such as those from Mixamo (Inc. 2021) and AMASS (Mahmood et al. 2019), remains largely untapped. This brings up the

*Work done during an internship at Huawei
†Corresponding author

question we are investigating in this work, as illustrated in Figure 1. Can we learn reliable diffusion models for controllable motion generation based on the supervision of *noisy* and the *unannotated* motion sequences?

Fortunately, with the inherent denoising mechanism of diffusion models, we can answer this question with a simple yet effective solution that applies separate diffusion steps regarding the source of training motion data, referred to as **MotionMix**. To demonstrate our application and approach, we split each gold annotated motion dataset into two halves: the first half of the motions are injected with random-scale Gaussian noises (noisy half), and the second half is deprived of annotations (clean half). As in Figure 2, the diffusion model bases on the clean samples for diffusion steps in $[1, T^*]$, with condition input erased. Meanwhile, noisy motions supervise the model with explicit conditions for the rest of steps $[T^* + 1, T]$. Note T^* is an experimental hyperparameter, with its role analyzed in later ablation studies. Our key insight is that, during sampling, starting from Gaussian noises, the model first produces rough motion approximations with conditional guidance in the initial $T-T^*$ steps; afterward, these rough approximations are further refined by unconditional sampling in the last T^* steps. Yet learning with weak supervision signals, our proposed MotionMix empirically facilitates motion generation with higher quality than fully supervised models on multiple applications. Benefiting from the conciseness of design, MotionMix finds its place in many applications. In this work, we thoroughly examine the effectiveness and flexibility of the proposed approach through extensive experiments on benchmarks of text-to-motion, music-to-dance, and action-to-motion tasks.

Our main contributions can be summarized as follows:

• We present **MotionMix**, the first weakly-supervised approach for conditional diffusion models that utilizes both *noisy annotated* and *clean unannotated* motion sequences simultaneously.

• We demonstrate that by training with these two sources of data simultaneously, **MotionMix** can improve upon prior state-of-the-art motion diffusion models across various tasks and benchmarks, without any conflict.

• Our approach opens new avenues for addressing the scarcity of clean and annotated motion sequences, paving the way for scaling up future research by effectively harnessing available motion resources.

2 Related Work

2.1 Weakly-Supervised Learning

To tackle the limited availability of annotated data, researchers have been exploring the use of semi-supervised generative models, using both annotated and unannotated data (Kingma et al. 2014; Li et al. 2017; Lucic et al. 2019). However, the investigation of semi-supervised diffusion models remains limited (You et al. 2023), possibly due to the significant performance gap observed between conditional and unconditional diffusion models (Bao et al. 2022; Dhariwal and Nichol 2021; Tevet et al. 2022). Moreover, many state-of-the-art models, such as Stable Diffusion (Rombach et al. 2021), implicitly assume the availability of abundant annotated data for training (Chang, Koulieris, and Shum 2023; Kawar et al. 2023). This assumption poses a challenge when acquiring high-quality annotated data is expensive, particularly in the case of 3D human motion data.

Recent interest has emerged in developing data-efficient approaches for training conditional diffusion models with low-quality data (Daras et al. 2023; Kawar et al. 2023), or utilizing unsupervised (Tur et al. 2023), semi-supervised (You et al. 2023), self-supervised methods (Miao et al. 2023). These approaches have exhibited promising results across various domains and hold potential for future exploration of diffusion models when handling limited annotated data. However, in the domain of human motion generation, efforts toward these approaches have been even more limited. One related work, Make-An-Animation (Azadi et al. 2023), trains a diffusion model utilizing unannotated motions in a semi-supervised setting. In contrast, our work introduces a unique aspect by training with noisy annotated motion and clean unannotated motion.

2.2 Conditional Motion Generation

Over the years, human motion generation has been extensively studied using various signals, including prefix poses (Ruiz, Gall, and Moreno-Noguer 2018; Guo et al. 2022c; Petrovich, Black, and Varol 2021), action classes (Guo et al. 2020; Petrovich, Black, and Varol 2021; Cervantes et al. 2022), textual descriptions (Guo et al. 2022b; Petrovich, Black, and Varol 2022; Guo et al. 2022a; Ahuja and Morency 2019; Bhattacharya et al. 2021), or music (Li et al. 2020, 2021; Siyao et al. 2022; Gong et al. 2023). However, it is non-trivial for these methods to align the distributions of motion sequences and conditions such as natural languages or speech (Chen et al. 2022). Diffusion models resolve this problem using a dedicated multi-step gradual diffuse and denosing process(Ramesh et al. 2022a; Saharia et al. 2022; Ho et al. 2022). Recent advancements, such as MDM (Tevet et al. 2022), MotionDiffuse (Zhang et al. 2022), MLD (Chen et al. 2022), have demonstrated the ability of diffusion-based models to generate plausible human motion, guided by textual descriptions or action classes. In the music domain, EDGE (Tseng, Castellon, and Liu 2022) showcased high-quality dance generation in diverse music categories. Nevertheless, these works still rely on high-quality motion datasets with annotated guidance.

3 Method

3.1 Problem Formulation

Conditional motion generation involves generating high-quality and diverse human motion sequences based on a desired conditional input c. This input can take various forms, such as a textual description $w^{1:N}$ of N words (Guo et al. 2022b), an action class $a \in A$ (Guo et al. 2020), music audio m (Li et al. 2021), or even an empty condition $c = \emptyset$ (unconditional input) (Raab et al. 2022). Our goal is to train a diffusion model in a weakly-supervised manner, using both noisy motion sequences with conditional inputs $c = \{\emptyset, a, w, c\}$ (where $\emptyset$ is used when the classifier-free guidance (Ho and Salimans 2022) is applied) and clean motion sequences with

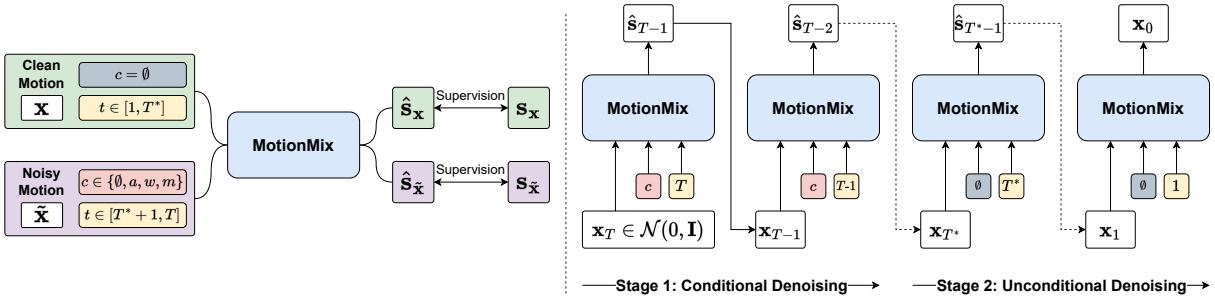

Figure 2: (Left) Training Process. The model is trained with a mixture of noisy and clean data. A noise timestep in ranges of $[1, T^*]$ and $[T^*+1, T]$ is sampled respectively for each clean and noisy data. Here, T^* is a denoising pivot that determines the starting point from which the diffusion model refines the noisy motion sequences into clean ones without any guidance. (Right) Sampling Process. The sampling process consists of two stages. In Stage-1 from timestep T to T^*+1, the model generates the rough motion approximations, guided by the conditional input c. In Stage-2 from timestep T^* to 1, the model refines these approximations to high-quality motion sequences while the input c is masked.

unconditional input $c = \emptyset$. Despite being trained with noisy motions, our model can consistently generate plausible motion sequences. To achieve this, we propose a two-stage reverse process, as illustrated in Figure 2.

3.2 Diffusion Probabilistic Model

The general idea of a diffusion model, as defined by the denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020), is to design a *diffusion process* that gradually adds noise to a data sample and trains a neural model to learn a *reverse process* of denoising it back to a clean sample. Specifically, the diffusion process can be modeled as a Markov noising process with $\{\mathbf{x}_t\}_{t=0}^{T}$ where $\mathbf{x}_0 \sim p(x)$ is the clean sample drawn from the data distribution. The noised $\mathbf{x}_t$ is obtained by applying Gaussian noise $\boldsymbol{\epsilon}_t$ to $\mathbf{x}_0$ through the posterior:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t \in (0, 1)$ are constants which follow a monotonically decreasing scheduler. Thus, when $\bar{\alpha}_t$ is small enough, we can approximate $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$.

In the reverse process, given the condition c, a neural model f_θ is trained to estimate the clean sample $\mathbf{x}_0$ (Ramesh et al. 2022b) or the added noise ϵ_t (Ho, Jain, and Abbeel 2020) for all t. The model parameters θ are optimized using the "simple" objective introduced by Ho, Jain, and Abbeel:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t\sim[1,T],\mathbf{s}_t}\left[\|\mathbf{s}_t - f_\theta(\mathbf{x}_t, t, c)\|^2\right] \tag{2}$$

where the target objective $\mathbf{s}_t$ refers to either $\mathbf{x}_0$ or $\boldsymbol{\epsilon}_t$ for ease of notation.

3.3 Training

We propose a novel weakly-supervised learning approach that enables a diffusion model to effectively utilize both noisy and clean motion sequences. During the training phase, we construct batches comprising both noisy and clean samples, each coupled with a corresponding guidance condition c, as further detailed in Subsection 3.5. To learn the denoising process, we apply the diffusion process to this batch using Equation 1 with varying noise timesteps. In contrary to the conventional training, where both noisy and clean motion sequences are treated as the ground truth x_0 with diffusion steps spanning $[1, T]$, our approach adopts separate ranges for different data types. For noisy samples, we randomly select noise timesteps $t \in [T^*+1, T]$, while for clean samples, we confine them to $t \in [1, T^*]$. Here, T^* serves as a denoising pivot, determining when the diffusion model starts refining noisy motion sequences into cleaner versions. This pivot is especially crucial in real-world applications, where motion capture data might be corrupted by noise due to diverse factors. This denoising strategy for *noisy motions* draws inspiration from (Nie et al. 2022), which purified adversarial images by diffusing them up to a specific timestep T^* before denoising to clean images. The determination of T^* typically relies on empirical estimation, its impact on generation quality is further analyzed in Table 4.

Through this training process, the model becomes adept at generating initial rough motions from T to T^*+1, and subsequently refining these rough motions into high-quality ones from T^* to 1. By dividing into two distinct time ranges, the model can effectively learn from both noisy and clean motion sequences as ground truth without any conflict.

3.4 Two-stage Sampling and Guidance

Our approach introduces a modification to the conventional DDPM sampling procedure, which commonly relies on the same explicit conditional input c to guide the denoising operation at each time step t, initiating from T and denoising back to the subsequent time step $t-1$ until reaching $t = 0$. However, it is important to note that our work specifically focuses on clean, unannotated samples. As discussed in Subsection 3.3, these samples are trained using an identical guidance condition $c = \emptyset$ confined within the time interval $[1, T^*]$. Consequently, if the conventional DDPM sampling process is employed within this temporal range, it could potentially lead to jittering or the generation of unrealistic motions. This occurs because the model is not trained to handle varying conditions within this specific range. To tackle this issue, we adopt a distinct strategy to align the sampling pro-

cess accordingly. Specifically, when the model reaches the denoising pivot T^* during the sampling, we substitute the conditional input with $c = \emptyset$ starting from T^*.

In the case of using classifier-free guidance (Ho and Salimans 2022), guided inference is employed for all t, which involves generating motion samples through a weighted sum of unconditionally and conditionally generated samples:

$$\hat{\mathbf{s}}(\mathbf{x}_t, t, c) = w \cdot f_\theta(\mathbf{x}_t, t, c) + (1 - w) \cdot f_\theta(\mathbf{x}_t, t, \emptyset) \quad (3)$$

where w is the guidance weight during sampling.

3.5 Data Preparation

To facilitate our setting, we randomly partition an existing training dataset into two subsets. In one subset, we retain the annotated condition and introduce noise to the motion sequences to approximate the real noisy samples. In the other subset, we reserve the cleanliness of the data and discard the annotated conditions by replacing them as $c = \emptyset$.

Motivated by the use of Gaussian noises in approximating noisy samples in prior works (Tiwari et al. 2022; Fiche et al. 2023), we apply Equation 1 to gradually introduce noise to the clean samples. Since the precise noise schedule in real-world motion capture data is unknown, we address this uncertainty by applying a random noising step sampled from the range $[T_1, T_2]$, where T_1 and T_2 are hyperparameters simulating the level of disruption in real noisy motions. Interestingly, our experiments (Tab. 6) show that neither smaller value of T_1, T_2 nor small T_2-T_1 relates to better performance. Due to page limit, examples of noisy motions for training are presented in supplementary videos.

It is worth noting that the processes of dividing the training dataset and preparing noisy samples, and unannotated samples only take place on the side of the training dataset. The remaining evaluation dataset, diffusion models, and training process are kept unchanged as in previous works.

4 Experiments

We thoroughly experiment our MotionMix in diverse tasks using different conditional motion generation diffusion models as backbones: **(1)** MDM (Tevet et al. 2022) for text-to-motion task on HumanML3D (Guo et al. 2022b), KIT-ML (Plappert, Mandery, and Asfour 2016), as well as action-to-motion task on HumanAct12 (Guo et al. 2020) and UESTC (Ji et al. 2018); **(2)** MotionDiffuse (Zhang et al. 2022) for text-to-motion task; and **(3)** EDGE (Tseng, Castellon, and Liu 2022) for music-to-dance task on AIST++ (Li et al. 2021). For details of each benchmark and model, please refer to Appendices A and B, respectively.

4.1 Text-to-motion

• **Implementation Details.** On both datasets, we train the MDM and MotionDiffuse models from scratch for $700K$ and $200K$ steps, respectively. To approximate the noisy motion data $\tilde{\mathbf{x}}$ from $\mathbf{x} \in \mathbb{R}^{N \times D}$, we use noisy ranges $[20, 60]$ and $[20, 40]$ for HumanML3D and KIT-ML, respectively.

• **Evaluation Metrics.** As suggested by Guo et al., the metrics are based on a text feature extractor and a motion feature extractor jointly trained under contrastive loss to produce feature vectors for matched text-motion pairs. R Precision (top 3) measures the accuracy of the top 3 retrieved descriptions for each generated motion, while the Frechet Inception Distance (FID) is calculated using the motion extractor as the evaluator network. Multimodal Distance measures the average Euclidean distance between the motion feature of each generated motion and the text feature of its corresponding description in the test set. Diversity measures the variance of the generated motions across all action categories, while MultiModality measures the diversity of generated motions within each condition.

• **Quantitative Result.** Table 1 presents quantitative results of our weakly-supervised MotionMix using MDM and MotionDiffuse backbones, in comparison with their original models that are trained with fully annotated and clean motion sequences. To our surprise, in most settings, MotionMix even improves the motion quality (i.e., FID) and multimodal consistency (i.e., R Precision) upon the fully supervised backbones. For example, on HumanML3D and KIT-ML dataset, MDM (**MotionMix**) commonly reduces FID by over 0.16 compare to MDM; this comes with the enhancement of both R Precision and Multimodal Distance. We may attribute this to the better generalizability and robustness by involving noisy data in our MotionMix. On the specifical setting of MotionDiffuse (**MotionMix**) on HumanML3D, though being inferior to the original MotionDiffuse, our MotionMix maintains competitive performance on par with other fully supervised baselines, such as Language2Pose (Ahuja and Morency 2019), Text2Gestures (Bhattacharya et al. 2021), Guo et al. (Guo et al. 2022b).

4.2 Action-to-motion

• **Implementation Details.** Following the experimental setup by Tevet et al., we train the MDM (**MotionMix**) from scratch on the HumanAct12 and UESTC datasets for $750K$ and $2M$ steps, respectively. In our approximation preprocess, we determine the amount of noise to be injected into both the pose sequence $\mathbf{p}$ and the root translation $\mathbf{r}$ by randomly sampling from range $[10, 30]$. The resulting $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{r}}$ are then concatenated to obtain noisy motion $\tilde{\mathbf{x}}$.

• **Evaluation Metrics.** Four metrics are used to assess the quality of generated motions. The FID is commonly used to evaluates the overall quality of generated motions. Accuracy (Acc.) measures the correlation between the generated motion and its action class. Diversity (Div.) and MultiModality (MM) are similar to the text-to-motion metrics.

• **Quantitative Result.** Table 2 presents the performance outcomes of MDM (**MotionMix**) and several baseline models, including Action2Motion (A2M) (Guo et al. 2020), ACTOR (Petrovich, Black, and Varol 2021), INR (Cervantes et al. 2022), MLD (Chen et al. 2022), and MDM (Tevet et al. 2022), on both the HumanAct12 and UESTC datasets. Following the methodology of Tevet et al., we perform 20 evaluations, each comprising 1000 samples, and present average scores with a confidence interval of 95%. The results highlight that our MotionMix achieves compet-

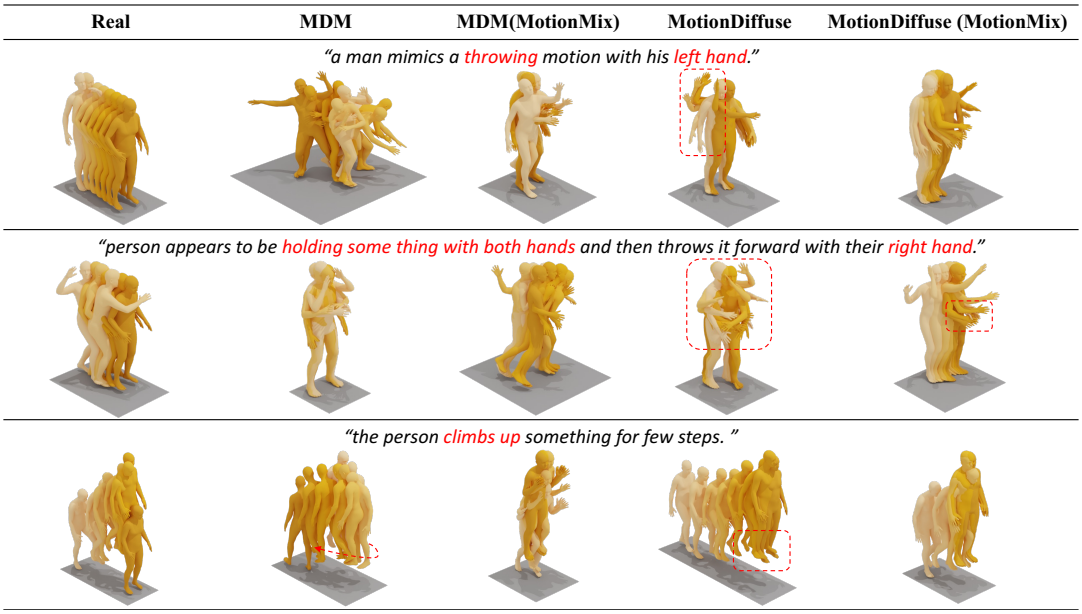

Figure 3: Qualitative performance of baseline MDM and MotionDiffuse models, trained exclusively on high-quality annotated data, with our MotionMix approach, which learns from imperfect data sources. Their visualized motion results are presented alongside real references for three distinct text prompts. Please refer to supplementary files for more animations.

itive performance with significantly fewer high-quality annotated data instances. In particular, the improvement seen on the UESTC dataset underscores its efficacy in training with noisy motion data from the real-world scenario. On the other hand, the deterioration in performance on HumanAct12 suggests that our approach is better suited for larger datasets, given that the size of HumanAct12 is remarkably smaller than that of UESTC. Nevertheless, our supplementary videos demonstrate that the model trained on HumanAct12 remains capable of generating quality motion sequences based on the provided action classes.

4.3 Music-to-dance

• **Implementation Details.** Similar to the action-to-motion task, we inject noise into both $\mathbf{p}$ and $\mathbf{r}$ using the same noise timestep sampled from $[20, 80]$. Since the contact label $\mathbf{b}$ is obtained from both $\mathbf{p}$ and $\mathbf{r}$, it is not necessary to inject noise into $\mathbf{b}$. Following the setup of Tseng, Castellon, and Liu, we train both the EDGE model and our EDGE (**MotionMix**) from scratch on AIST++ for 2000 epochs.

• **Evaluation Metrics.** We adopt the evaluation settings from the paper EDGE, including Physical Foot Contact (PFC), Beat Alignment, and Diversity metrics. PFC gauges physical plausibility by capturing realistic foot-ground contact, without explicit modeling or static contact assumptions. Beat Alignment assesses the synchronization of dances with music beats, following prior works (Li et al. 2021; Siyao et al. 2022). Diversity is measured in both kinetic (Dist_k) and geometric (Dist_g) feature spaces.

• **Quantitative Result.** In contrary to prior works, which typically reported only a single evaluation result, we have observed that the metrics can be inconsistent. Thus, to offer a more comprehensive evaluation, we run the evaluation 20 times similar to the previous two tasks for our retrained EDGE model and our EDGE (**MotionMix**) variant. For Bailando (Siyao et al. 2022) and FACT (Li et al. 2021), we directly fetched results from the paper EDGE (Tseng, Castellon, and Liu 2022). The results in Table 3 vividly demonstrate that, our EDGE (**MotionMix**) significantly outperforms the baseline across all metrics, showcasing improvements of up to 43.1% in PFC and 95.0% in Dist_k. This further reinforces the generalizability prowess of MotionMix, consistent with the outcomes observed in text-to-motion experiments.

5 Ablation Studies

MotionMix is introduced as a potential solution that enables the diffusion model to effectively leverage both noisy motion sequences and unannotated data. To demonstrate the efficacy of this approach, we approximate noisy samples from existing datasets and train the model on them, which incorporate several essential hyperparameters: (1) the denoising pivot T^*; (2) the ratio of noisy and clean data for training; (3) the noisy range $[T_1, T_2]$ to approximate noisy data. In this section, we thoroughly assess the impact of each hyperparameters within MotionMix. All ablation experiments are carried out on the HumanML3D dataset using the MDM model with the same settings described in Subsection 4.1.

5.1 Effect of The Denoising Pivot T^*

We begin our ablation studies by examining the impact of the denoising pivot T^*. To evaluate its impact, we conduct experiments with a fixed noisy range of $[T_1, T_2] = [20, 60]$, a noisy ratio of 50%, and evaluate various T^* values, encom-

	Method	R Precision (top 3)↑	FID↓	Multimodal Dist.↓	Diversity→	Multimodality↑
HumanML3D	Real Motion	0.797	0.002	2.974	9.503	-
	Language2Pose	0.486	11.02	5.296	7.676	-
	Text2Gestures	0.345	7.664	6.030	6.409	-
	Guo et al.	0.740	1.067	3.340	9.188	2.090
	MLD	0.772	0.473	3.196	9.724	2.413
	MDM	0.611	0.544	5.566	9.559	2.799
	MDM†	0.632 (↑3.4%)	0.381 (↑30.0%)	5.325 (↑4.3%)	9.520 (↑69.6%)	2.718 (↓2.9%)
	MotionDiffuse	0.782	0.630	3.113	9.410	1.553
	MotionDiffuse†	0.738 (↓5.6%)	1.021 (↓62.1%)	3.310 (↓6.3%)	9.297 (↓121.5%)	1.523 (↓1.9%)
KIT-ML	Real Motion	0.779	0.031	2.788	11.080	-
	Language2Pose	0.483	6.545	5.147	9.073	-
	Text2Gestures	0.338	12.12	6.964	9.334	-
	Guo et al.	0.693	2.770	3.401	10.910	1.482
	MLD	0.734	0.404	3.204	10.800	2.192
	MDM	0.396	0.497	9.191	10.847	1.907
	MDM†	0.404 (↑2.0%)	0.322 (↑35.2%)	9.068 (↑1.3%)	10.781 (↓28.3%)	1.946 (↑2.0%)
	MotionDiffuse	0.739	1.954	2.958	11.100	0.730
	MotionDiffuse†	0.742 (↑0.4%)	1.192 (↑39.0%)	3.066 (↓3.6%)	10.998 (↓310%)	1.391 (↑90.5%)

Table 1: Quantitative results of text-to-motion on the test set of HumanML3D and KIT-ML. Note all baselines are trained with gold data. † means the <model> is trained in our weakly-supervised setting. We run all the evaluation 20 times (except *Multimodality* runs 5 times) and reports only the mean score due to limited space. ↑ means higher is better, ↓ means lower is better, → means closer to the real distribution is better. The ↑ $x\%$ and ↓ $x\%$ indicate the percentage difference in performance improvement or deterioration when comparing our approach to its correspond baseline.

	Method	FID ↓	Acc. ↑	Div. →	MM →
HumanAct12	Real Motion	0.053	0.995	6.835	2.604
	A2M	0.338	0.917	6.850	2.511
	ACTOR	0.120	0.955	6.840	2.530
	INR	0.088	0.973	6.881	2.569
	MLD	0.077	0.964	6.831	2.824
	MDM	0.100	0.990	6.860	2.520
	MDM†	0.196 (↓96%)	0.930 (↓6.1%)	6.836 (↑96%)	3.043 (↓423%)
UESTC	Real Motion	2.790	0.988	33.349	14.160
	ACTOR	23.430	0.911	31.960	14.520
	INR	15.000	0.941	31.590	14.680
	MLD	15.790	0.954	33.520	13.570
	MDM	12.810	0.950	33.100	14.260
	MDM†	11.400 (↑11%)	0.960 (↑1.1%)	32.806 (↓118%)	14.277 (↓17%)

Table 2: Quantitative results of action-to-motion on the HumanAct12 dataset and UESTC test set. We run the evaluation 20 times, the metric details and † are similar to Table 1.

Method	PFC ↓	Beat Align. ↑	Dist_k →	Dist_g →
Real Motion	1.380	0.314	9.545	7.766
Bailando	1.754	0.23	10.58	7.72
FACT	2.2543	0.22	10.85	6.14
EDGE‡	3.494	0.226	20.684	9.145
EDGE†	1.988 (↑43.1%)	0.256 (↑13.3%)	10.103 (↑95.0%)	6.595 (↑15.1%)

Table 3: Quantitative results of music-to-dance on the AIST++ test set. We run the evaluation 20 times, the metric details and † are similar to Table 1. ‡ denotes the EDGE model that is re-trained by us.

passing 20, 40, 60, and 80. The results, detailed in Table 4, reveal a notable observation: a roughly estimated denoising pivot is sufficient for real-world scenarios, as evidenced by the competitive outcomes across various T^* values. This robustness underlines the versatility of our MotionMix approach. Additionally, selecting a very small denoising pivot (e.g., $T^* = 0$ or 20) enables conditions to steer the model toward diverse rough motion sequences before the refining phase, as reflected in the MModality score trend. However, this small value may potentially compromise motion quality, leading to subpar results in other metrics. In contrast, the choice of $T^* = 60$, which is well aligned with our predefined noisy range, yields superior results in multiple evaluation metrics. This sheds light on the need of tuning the denoising pivot to optimize the results, as this hyperparameter determines the starting point for the diffusion model to transform initial noisy motion into high-quality sequences.

Method	R Prec. (top 3)↑	FID↓	Multi. Dist.↓	Div.→	MM↑
Real Motion	0.797	0.00	2.97	9.50	-
MDM	<u>0.611</u>	0.54	5.57	<u>9.56</u>	2.79
50% noisy,T_1=20,T_2=60					
MDM† *(T^*=0)*	0.598	0.71	<u>5.50</u>	9.75	**3.04**
MDM† *(T^*=20)*	0.601	0.50	5.56	9.41	<u>2.94</u>
MDM† *(T^*=40)*	0.604	<u>0.40</u>	5.52	9.40	2.75
MDM† *(T^*=60)*	**0.632**	**0.38**	**5.33**	**9.52**	2.72
MDM† *(T^*=80)*	0.594	0.59	5.67	9.24	2.60

Table 4: We evaluate MDM (MotionMix) on HumanML3D test set using different values of the denoising pivot T^*. The metrics details are similar to Table 1. The best and the second best result are bold and underlined respectively.

Method	R Prec. (top 3)↑	FID↓	Multi. Dist.↓	Div.→	MM↑
Real Motion	0.797	0.00	2.97	9.50	-
MDM	0.611	0.54	5.57	9.56	2.79
T_1=20,T_2=60,T^*=60					
MDM† *(30% noisy)*	0.601	0.90	5.58	9.08	<u>2.86</u>
MDM† *(50% noisy)*	**0.632**	<u>0.38</u>	**5.33**	**9.52**	2.72
MDM† *(70% noisy)*	<u>0.615</u>	**0.36**	<u>5.55</u>	<u>9.46</u>	**2.87**

Table 5: We evaluate MDM (MotionMix) on HumanML3D test set using different ratios for noisy and clean data. The metrics details are similar to Table 1. The best and the second best result are bold and underlined respectively.

5.2 Effect of Noisy/Clean Data Ratio

We evaluate how the noisy/clean data ratio affects MotionMix by keeping $T^* = 60$ and $[T_1, T_2] = [20, 60]$ constant. We experiment with various noisy ratios of 30%, 50%, and 70%. The results in Table 5 show interesting trends across the evaluation metrics. Notably, higher noisy ratios (i.e., 50% and 70%) consistently outperform the lower ratio (i.e., 30%). Note that a higher noisy ratio allows the model to access more annotated text conditions, yielding better R Precision and Multimodal Distance. On the other hand, the 30% ratio, despite being trained with a greater amount of clean data, exhibits suboptimal motion quality (scoring 0.90 in FID) in comparison to other supervised baselines in Table 1, such as Language2Pose (FID of 11.02), Text2Gestures (FID of 7.66), Guo et al. (FID of 1.067). Nevertheless, it still achieves results on par with the supervised MDM baseline in terms of multimodal consistency (i.e. Multimodal Distance). These observations underscore the resilience of our MotionMix approach to variations in the noisy/clean data ratio.

5.3 Effect of The Noisy Range

The purpose of the noisy range in our work is to approximate the noise schedule found in real-world motion capture data. Thus, for different datasets in Section 4, we choose noisy ranges based on the visualization of motion from each dataset. For example, UESTC (Ji et al. 2018) contains noisy mocap data, while HumanML3D (Guo et al. 2022b), derived from AMASS (Mahmood et al. 2019), consists of clean motion sequences. This ablation, therefore, comprehensively evaluates the effectiveness of our MotionMix approach when handling different noisy levels of motion sequences. We categorize the evaluations into two groups: narrow/wide ranges of noise and low/high schedules of noise. All experiments are conducted with a noisy ratio of 50%, and the denoising pivot T^* is equal to the chosen T_2. The results are presented in Table 6.

Method	R Prec. (top 3)↑	FID↓	Multi. Dist.↓	Div.→	MM↑
50% noisy, $T^* = T_2$					
MDM† *(T_1=20,T_2=40)*	<u>0.616</u>	<u>0.45</u>	<u>5.46</u>	9.59	2.59
MDM† *(T_1=20,T_2=60)*	**0.632**	**0.38**	**5.33**	**9.52**	<u>2.72</u>
MDM† *(T_1=20,T_2=80)*	0.604	0.61	5.54	<u>9.55</u>	**2.77**
50% noisy, $T^* = T_2$					
MDM† *(T_1=10,T_2=30)*	0.592	0.71	5.63	9.57	2.78
MDM† *(T_1=20,T_2=40)*	**0.616**	<u>0.45</u>	**5.46**	9.59	2.59
MDM† *(T_1=40,T_2=60)*	<u>0.598</u>	0.55	5.60	**9.48**	<u>2.82</u>
MDM† *(T_1=60,T_2=80)*	0.597	**0.44**	<u>5.55</u>	<u>9.45</u>	**2.90**

Table 6: We evaluate MDM (MotionMix) on HumanML3D test set using different noisy ranges $[T_1, T_2]$ to approximate the noisy motion sequences. The table presents two distinct scenarios: the upper block ablates *how much the range spans*, while the lower block examines the impact of the *corruption level* of noisy motions. The metrics details are similar to Table 1. For each setting, the best and the second best result are bold and underlined respectively.

- **Narrow/Wide Noisy Range.** Three noisy ranges $[T_1, T_2] \in \{[20, 40], [20, 60], [20, 80]\}$ are set to analyze the effect of *how much the range spans*. Counterintuitively, a smaller noisy range does not equal better performance. For example, noisy range $[20, 60]$ leads to overall the best performance, compared to $[20, 40]$. Though, larger noisy range (i.e., $[20, 80]$) unevitably deteriotate the model capacity.
- **Low/High Noisy Schedule.** Four contrast ranges $[T_1, T_2] \in \{[10, 30], [20, 40], [40, 60], [60, 80]\}$ are set to evaluate the robustness of MotionMix regarding *corruption level* of noisy motions. Notably, MotionMix performs reasonably stable on different levels of corrupted motions.

6 Conclusion

This work addresses the training challenge of a conditional human motion generation model with both *noisy annotated* and *clean unannotated* motion sequences. Our proposed approach, **MotionMix**, pioneers the utilization of a weakly-supervised diffusion model to tackle this challenge. MotionMix effectively overcomes constraints posed by limited high-quality annotated data, demonstrating competitive results against fully-supervised models. The versatility of MotionMix is evident across various tasks/benchmarks and fundamental diffusion model designs. Comprehensive ablation studies further bolster its resilience in diverse noisy schedules and the strategic selection of the denoising pivot.

References

Ahuja, C.; and Morency, L.-P. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. *2019 International Conference on 3D Vision (3DV)*, 719–728.

Aristidou, A.; Yiannakidis, A.; Aberman, K.; Cohen-Or, D.; Shamir, A.; and Chrysanthou, Y. 2021. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *IEEE transactions on visualization and computer graphics*, PP.

Azadi, S.; Shah, A.; Hayes, T.; Parikh, D.; and Gupta, S. 2023. Make-An-Animation: Large-Scale Text-conditional 3D Human Motion Generation. *ArXiv*, abs/2305.09662.

Bao, F.; Li, C.; Sun, J.; and Zhu, J. 2022. Why Are Conditional Generative Models Better Than Unconditional Ones? *ArXiv*, abs/2212.00362.

Bhattacharya, U.; Rewkowski, N.; Banerjee, A.; Guhan, P.; Bera, A.; and Manocha, D. 2021. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents. *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 1–10.

Cervantes, P.; Sekikawa, Y.; Sato, I.; and Shinoda, K. 2022. Implicit Neural Representations for Variable Length Human Motion Generation. *ArXiv*, abs/2203.13694.

Chang, Z.; Koulieris, G. A.; and Shum, H. P. H. 2023. On the Design Fundamentals of Diffusion Models: A Survey. *ArXiv*, abs/2306.04542.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; Yu, J.; and Yu, G. 2022. Executing your Commands via Motion Diffusion in Latent Space. *ArXiv*, abs/2212.04048.

Choutas, V.; Pavlakos, G.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2020. Monocular Expressive Body Regression through Body-Driven Attention. *ArXiv*, abs/2008.09062.

Daras, G.; Shah, K.; Dagan, Y.; Gollakota, A.; Dimakis, A. G.; and Klivans, A. R. 2023. Ambient Diffusion: Learning Clean Distributions from Corrupted Data. *ArXiv*, abs/2305.19256.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *ArXiv*, abs/2105.05233.

Fiche, G.; Leglaive, S.; Alameda-Pineda, X.; and S'eguier, R. 2023. Motion-DVAE: Unsupervised learning for fast human motion denoising. *ArXiv*, abs/2306.05846.

Gong, K.; Lian, D.; Chang, H.; Guo, C.; Zuo, X.; Jiang, Z.; and Wang, X. 2023. TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration. arXiv:2304.02419.

Guo, C.; Xuo, X.; Wang, S.; and Cheng, L. 2022a. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. *ArXiv*, abs/2207.01696.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022b. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. *Proceedings of the 28th ACM International Conference on Multimedia*.

Guo, W.; Du, Y.; Shen, X.; Lepetit, V.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2022c. Back to MLP: A Simple Baseline for Human Motion Prediction. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4798–4808.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A. A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *ArXiv*, abs/2210.02303.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *ArXiv*, abs/2006.11239.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598.

Inc., A. S. 2021. Mixamo. https://www.mixamo.com/. Accessed: 2021-12-25.

Ji, Y.; Xu, F.; Yang, Y.; Shen, F.; Shen, H. T.; and Zheng, W. 2018. A Large-scale RGB-D Database for Arbitrary-view Human Action Recognition. *Proceedings of the 26th ACM international conference on Multimedia*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2017. End-to-End Recovery of Human Shape and Pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7122–7131.

Kawar, B.; Elata, N.; Michaeli, T.; and Elad, M. 2023. GSURE-Based Diffusion Model Training with Corrupted Data. *ArXiv*, abs/2305.13128.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised Learning with Deep Generative Models. *ArXiv*, abs/1406.5298.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2019. VIBE: Video Inference for Human Body Pose and Shape Estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5252–5262.

Koppula, H. S.; and Saxena, A. 2013. Anticipating human activities for reactive robotic response. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2071–2071.

Lee, L.-H.; Braud, T.; Zhou, P.; Wang, L.; Xu, D.; Lin, Z.; Kumar, A.; Bermejo, C.; and Hui, P. 2021. All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda. *ArXiv*, abs/2110.05352.

Li, C.; Xu, T.; Zhu, J.; and Zhang, B. 2017. Triple Generative Adversarial Nets. *ArXiv*, abs/1703.02291.

Li, J.; Yin, Y.; Chu, H.; Zhou, Y.; Wang, T.; Fidler, S.; and Li, H. 2020. Learning to Generate Diverse Dance Motions with Transformer. *ArXiv*, abs/2008.08171.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13381–13392.

Ling, H. Y.; Zinno, F.; Cheng, G.; and van de Panne, M. 2020. Character controllers using motion VAEs. *ACM Transactions on Graphics (TOG)*, 39: 40:1 – 40:12.

Lucic, M.; Tschannen, M.; Ritter, M.; Zhai, X.; Bachem, O.; and Gelly, S. 2019. High-Fidelity Image Generation With Fewer Labels. *ArXiv*, abs/1903.02271.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture As Surface Shapes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5441–5450.

Miao, Y.-C.; Zhang, L.; Zhang, L.; and Tao, D. 2023. DDS2M: Self-Supervised Denoising Diffusion Spatio-Spectral Model for Hyperspectral Image Restoration. *ArXiv*, abs/2303.06682.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. arXiv:2205.07460.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10965–10975.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. *ArXiv*, abs/2204.14109.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT Motion-Language Dataset. *Big Data*, 4(4): 236–252.

Raab, S.; Leibovitch, I.; Li, P.; Aberman, K.; Sorkine-Hornung, O.; and Cohen-Or, D. 2022. MoDi: Unconditional Motion Synthesis from Diverse Data. arXiv:2206.08010.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022a. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022b. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Ruiz, A. H.; Gall, J.; and Moreno-Noguer, F. 2018. Human Motion Prediction via Spatio-Temporal Inpainting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7133–7142.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv*, abs/2205.11487.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11040–11049.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human Motion Diffusion Model. *ArXiv*, abs/2209.14916.

Tiwari, G.; Antic, D.; Lenssen, J. E.; Sarafianos, N.; Tung, T.; and Pons-Moll, G. 2022. Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields. In *European Conference on Computer Vision (ECCV)*.

Tseng, J.-H.; Castellon, R.; and Liu, C. K. 2022. EDGE: Editable Dance Generation From Music. *ArXiv*, abs/2211.10658.

Tur, A. O.; Dall'Asen, N.; Beyan, C.; and Ricci, E. 2023. Exploring Diffusion Models for Unsupervised Video Anomaly Detection. *ArXiv*, abs/2304.05841.

You, Z.; Zhong, Y.; Bao, F.; Sun, J.; Li, C.; and Zhu, J. 2023. Diffusion Models and Semi-Supervised Learners Benefit Mutually with Few Labels. *ArXiv*, abs/2302.10586.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *ArXiv*, abs/2208.15001.