This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PersonaBooth: Personalized Text-to-Motion Generation

Boeun Kim[1,2,3] Hea In Jeong[2] JungHoon Sung[3] Yihua Cheng[1] Jeongmin Lee[2]
Ju Yong Chang[4] Sang-Il Choi[3] Younggeun Choi[3] Saim Shin[2] Jungho Kim[2] Hyung Jin Chang[1]

[1]University of Birmingham [2]Korea Electronics Technology Institute [3]Dankook University [4]Kwangwoon University
http://boeun-kim.github.io/page-PersonaBooth

Abstract

This paper introduces Motion Personalization, a new task that generates personalized motions aligned with text descriptions using several basic motions containing Persona. To support this novel task, we introduce a new large-scale motion dataset called ***PerMo (PersonaMotion)****, which captures the unique personas of multiple actors. We also propose a multi-modal finetuning method of a pretrained motion diffusion model called* ***PersonaBooth****. PersonaBooth addresses two main challenges: i) A significant distribution gap between the persona-focused PerMo dataset and the pretraining datasets, which lack persona-specific data, and ii) the difficulty of capturing a consistent persona from the motions vary in content (action type). To tackle the dataset distribution gap, we introduce a persona token to accept new persona features and perform multi-modal adaptation for both text and visuals during finetuning. To capture a consistent persona, we incorporate a contrastive learning technique to enhance intra-cohesion among samples with the same persona. Furthermore, we introduce a context-aware fusion mechanism to maximize the integration of persona cues from multiple input motions. PersonaBooth outperforms state-of-the-art motion style transfer methods, establishing a new benchmark for motion personalization.*

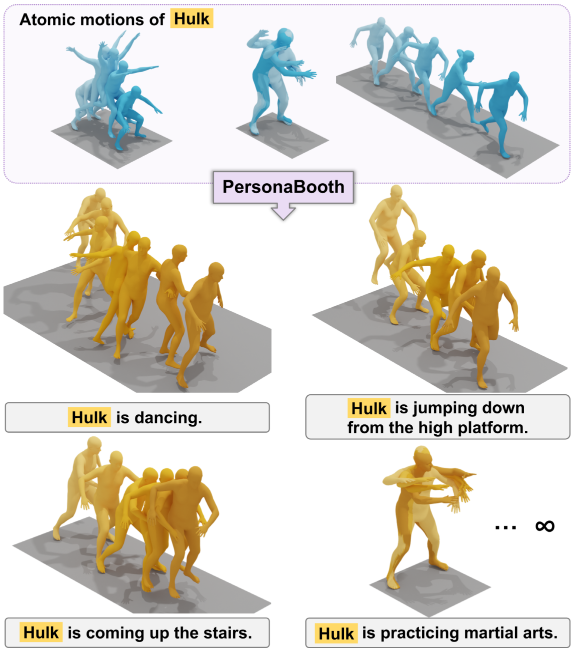

Atomic motions of Hulk
PersonaBooth
Hulk is dancing.
Hulk is jumping down from the high platform.
Hulk is coming up the stairs.
Hulk is practicing martial arts.
... ∞

Figure 1. Motion Personalization generates text-driven, personalized motions based on persona embedded in atomic input motions. We propose a new framework, **PersonaBooth**, along with a new benchmark dataset, **PerMo**, for Motion Personalization

1. Introduction

Imagine a scenario where, by recording just a few basic movements of yourself, an avatar in a virtual world can mirror your personal traits and unique motion style. This kind of 'motion personalization' enables realistic interactions in virtual spaces such as games and the metaverse [14]. Moreover, if avatar motions could be directed through text, creating video content would become easy without the need for real-life stunt actors [2, 24].

In this paper, we propose a task called *Motion Personalization*, which aims to generate text-based motions that reflect individual personas using a few key atomic movements, such as jumping, punching, or walking. We define a persona as the unique style expression of an individual. As illustrated in Fig. 1, this task aims to generate a broad range of realistic motions robustly. To facilitate this, we introduce a model called **PersonaBooth** and a new benchmark, **PerMo (PersonaMotion)**, specifically designed for motion personalization. Table 1 provides a comparison of similar motion generation tasks. Text-to-Motion Generation approaches [6, 21] generate motion using only text as input, without any reference motion, and numerous diffusion-based methods [3, 33, 35] have been proposed for this task. Motion Style Transfer (MST) focuses on transferring style

Table 1. Comparison of motion generation tasks

Motion Task	Existing Datasets	Source Motions	Multi-modal Finetuning
T2M Generation [3, 33, 35]	[6] [23]	None	✗
Style Transfer [25, 32, 38]	[34][1][19]	Single	✗
Personalization (Ours)	None	Multiple	✓

from a single source motion. While several datasets and methods are available for MST [25, 32, 34, 38], the use of a single source motion restricts the ability to generate the broad and diverse motions seen in real life [12]. The proposed *Motion Personalization* has no existing prior approaches or datasets. Furthermore, our proposed method introduces multi-modal finetuning—a process neglected in existing diffusion-based MST methods.

PersonaBooth is a multi-modal finetuning method for pretrained text-to-motion diffusion models. The challenge in finetuning diffusion models for the motion domain, compared to the image domain [4, 15, 17, 30, 31, 37], lies in the relatively lower diversity of pretraining data. The most commonly used pretraining dataset, HumanML3D [6], contains only 15K samples, which is significantly smaller than the large-scale image datasets with 400M-650M samples [27, 28]. Additionally, HumanML3D includes few data reflecting personality traits, creating a substantial distribution gap when finetuning with the persona-focused PerMo dataset. MoMo [25] found that only 55 samples include styles for ‘run,’ ‘walk,’ ‘jump,’ and ‘dance’ locomotions in HumanML3D, and we could not find any persona-related samples. Existing diffusion-based MST methods [25, 38] consider only the visual features of the input motion and rely on a fixed text feature, which limits their ability to incorporate new data into the textual component. This inflexibility can lead to a phenomenon known as ‘forgetting’, where the pretrained model loses its initial capabilities when finetuned with new data [38]. To address this, we introduce learnable *persona tokens* to capture persona features from new data and propose an adaptation scheme for both text and visuals.

Another challenge is extracting a consistent persona across different atomic motions, which manifests in diverse ways depending on the action content. For example, the elegance of a ballerina is expressed in her feet as she walks and in her hands as she waves. MST methods have faced similar challenges in disrupting the content of the target motion when transferring styles from the source motion with different content. Approaches like those of Park *et al*. [20], Jang *et al*. [9], and MoMo [25] cite this as a limitation. In contrast, MoST [12] addresses the issue with a style disentanglement loss, while MCM-LDM [32] uses the target motion’s trajectory to preserve content. However, MoST’s loss function requires running the model twice for two different input motions, which is computationally expensive, especially for diffusion models with numerous steps. Additionally, since our goal is to generate motion from text rather than from target motions, the method used by MCM-LDM is not applicable. Therefore, we introduce a novel contrastive learning-based loss specifically designed for text-to-motion diffusion, called *persona cohesion loss*. This loss facilitates cohesion across motion features with different content but the same persona.

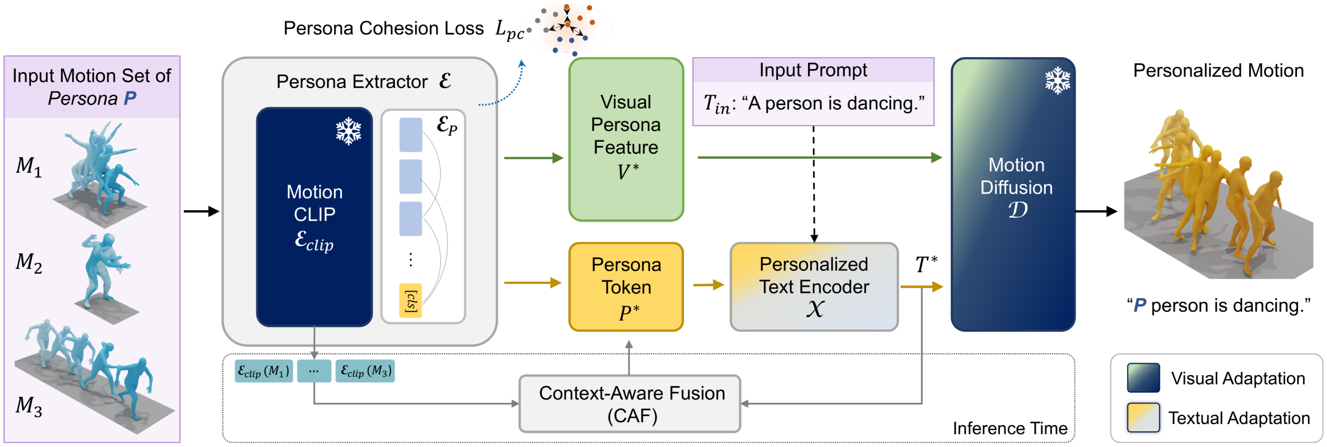
In addition, while MST uses a single input motion, we propose a new fusion method for multiple motions. In the image domain, InstantBooth [31] simply averages persona features, but this leads to unnatural blending in the motion. To address this, we introduce *Context-Aware Fusion (CAF)*, which assigns weights to the fusion of persona features based on the similarity between the prompt and input motions. The summarized contributions of this paper are:

- We introduce the Motion Personalization task and present PerMo, a new large-scale persona dataset captured by multiple actors.
- We propose PersonaBooth, a multi-modal finetuning method for diffusion models that introduces text adaptation through a persona token. PersonaBooth applies contrastive learning to effectively capture personas across various content types.
- PersonaBooth achieves strong performance in both Motion Personalization and MST tasks and maximizes performance through a CAF for multiple inputs.

2. Related Work

The Motion Style Transfer (MST) aims to transfer the style from a source motion to a target motion. The pioneering work by Aberman *et al*. [1] introduced the AdaIN layer, which simply replaces style statistics in specific layers during training. MotionPuzzle [9] proposes a Graph Convolution Network-based framework that transfers manually specified motion to a desired body part, enabling the combination of multiple source motions. MoST [12] addresses the issue of unwanted blending between different motion contents by introducing a transformer-based model [11] that more clearly separates style and content.

Recent MST algorithms have seen notable performance gains through the use of diffusion models. MoMo [25] proposes a zero-shot motion transfer method using a pretrained diffusion model, where it mixes two input motions within the attention module of the diffusion model. However, MoMo identifies a limitation: when the source motion content differs significantly from the target motion or text, the generated motion may not follow the target accurately. For example, applying the style of a stationary source motion to a walking target motion may still result in a stationary output. SMooDi [38] proposes a finetuning method for a pretrained diffusion model. However, due to the forgetting issue, it retrained the model with both the original pretrain-



Persona Cohesion Loss L_{pc}
Input Motion Set of Persona P
M_1
M_2
M_3
Persona Extractor $\mathcal{E}$
Motion CLIP $\mathcal{E}_{clip}$
$\mathcal{E}_P$
[cls]
Visual Persona Feature V^*
Persona Token P^*
Input Prompt
T_{in}: "A person is dancing."
Personalized Text Encoder $\mathcal{X}$
T^*
Motion Diffusion $\mathcal{D}$
Personalized Motion
"P person is dancing."
$\mathcal{E}_{clip}(M_1)$
...
$\mathcal{E}_{clip}(M_3)$
Context-Aware Fusion (CAF)
Inference Time
Visual Adaptation
Textual Adaptation

Figure 2. The overall framework of PersonaBooth. PersonaBooth has two adaptation paths—visual and text—for finetuning the Motion Diffusion model ($\mathcal{D}$). The Persona Extractor extracts both a visual persona feature (V^*) and a persona token (P^*) from the input motions. V^* is input into the adaptive layer of $\mathcal{D}$, while P^* is processed together with the input prompt through a Personalized Text Encoder, generating a personalized text feature, which is then input to $\mathcal{D}$. The entire model is trained with a classifier-free approach, incorporating a Persona Cohesion Loss. During inference, Context-Aware Fusion is applied for multiple input cases.

ing dataset and the style dataset. Additionally, the use of classifier-based training and the adaptation approach of creating a trainable copy of the entire model led to a training speed 10 times slower than the base diffusion model. MCM-LDM [32] addresses the challenges regarding two different contents of input motions by explicitly extracting the trajectory from the target motion and using it as a condition for the diffusion model. However, it faces a limitation: only trajectories similar to the target motion are generated, which reduces diversity. This contradicts our goal of creating diverse motions based on text prompts.

3. Methodology

We propose a method for multi-modal finetuning of a text-to-motion diffusion model, named PersonaBooth. The overall framework is shown in Fig. 2, and we aim to finetune the pretrained motion diffusion model, $\mathcal{D}$. The input includes multiple atomic motions $\{M_i\}$ and a text description T_{in}. PersonaBooth generates personalized motion that reflects the description. We use the same motion representation as HumanML3D [6], $M_i \in \mathbb{R}^{f \cdot 263}$, where f is the number of frames, and 263 includes joint rotations, positions, velocities, and foot contact. First, the Persona Extractor $\mathcal{E}$ processes each input motion to extract a visual persona feature V^* and a *persona token* P^*. Built on a pretrained motion clip model [22] ($\mathcal{E}_{clip}$), which captures the general characteristics of motion sequences, $\mathcal{E}$ includes a transformer structure ($\mathcal{E}_P$) specially designed to extract persona features. We applied a contrastive learning scheme to enhance the persona extraction capability of $\mathcal{E}_P$, which will be detailed in Sec. 3.1. The extracted V^* is then fed into a newly added adaptive layer in $\mathcal{D}$. Meanwhile, the Personalized Text Encoder $\mathcal{X}$ adaptively incorporates P^* to produce a personalized text feature T^*. T^* is input into $\mathcal{D}$, where $\mathcal{D}$ ultimately generates the personalized motion. During training, a single motion-text pair is provided to enable reconstruction. During inference, given motion sets $\{M_i\}$, the Context-Aware Fusion module generates weights to perform a weighted combination of the elements within $\{V_i^*\}$ and $\{P_i^*\}$. PersonaBooth explores efficient training by using a classifier-free guidance approach. During training, $\mathcal{E}_{clip}$ and $\mathcal{D}$, excluding the newly added adaptive layer, are kept frozen. Sec.3.1 to Sec.3.3 describe the process based on the training of single motion, M. The inference process for multiple input motions $\{M_i\}$ is explained in Sec.3.4.

3.1. Persona Extractor

Persona Extractor $\mathcal{E}$ extracts the *persona token* P^* along with the *visual persona feature* V^*. We used the TMR [22] structure for $\mathcal{E}_{clip}$, replacing their text encoder with the CLIP text encoder [26] and retraining it. The CLIP text encoder is the same one used in the Personalized Text Encoder $\mathcal{X}$, and therefore, this replacement enables P^* to share the same token embedding space as the input prompt. For input motion M, persona features V^* and P^* are extracted as

$$V^* = \mathcal{E}([cls], M]), \tag{1}$$

$$P^* = \text{MLP}(Y), \text{ where } Y = V^*[0], \tag{2}$$

where $[cls]$ denotes a class token.

To ensure the coherence of persona features across different motion contents of the same persona, we introduce the *persona cohesion loss* L_{pc}. We utilize a supervised contrastive learning scheme [10], where motion samples from the same persona are encouraged to be closer in the persona feature space, while motion samples from different personas are pushed apart. This helps to guide the persona fea-

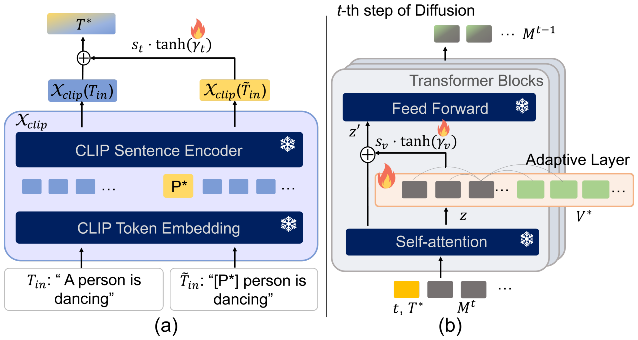

T^*
$s_t \cdot \tanh(\gamma_t)$
$\mathcal{X}_{clip}(T_{in})$
$\mathcal{X}_{clip}(\tilde{T}_{in})$
$\mathcal{X}_{clip}$
CLIP Sentence Encoder
P*
CLIP Token Embedding
T_{in}: " A person is dancing"
$\tilde{T}_{in}$: "[P*] person is dancing"
(a)
t-th step of Diffusion
M^{t-1}
Transformer Blocks
Feed Forward
z'
$s_v \cdot \tanh(\gamma_v)$
Adaptive Layer
z
V^*
Self-attention
t, T^*
M^t
(b)

Figure 3. Textual and visual adaptation. (a) Personalized Text Encoder, $\mathcal{X}$. (b) t-th step of the Motion Diffusion, $\mathcal{D}$. Learnable parameters are denoted by the fire icon

tures to form well-defined clusters for each persona. For a batch containing N input motion data, L_{pc} is applied to a total of $2N$ data points, with one positive sample for each input motion. The positive sample is randomly selected from the same persona group as the input motion. L_{pc} for a positive pair of index (i, j) is defined as

$$L_{pc} = -\log \frac{\exp(\text{sim}(h(Y_i), h(Y_j))/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(h(Y_i), h(Y_k))/\tau)}, \quad (3)$$

where $h(\cdot)$ denotes projection head. Y_j is extracted from M_j which belongs to the same persona group as M. $\text{sim}(u, v)$ indicates the cosine similarity between two vectors u and v. τ denotes the temperature parameter.

3.2. Textual and Visual Adaptation

The discrepancy between our PerMo dataset and the non-persona-specific pretraining dataset, HumanML3D [6] necessitates a carefully designed adaptation mechanism for effective finetuning. To address this, we propose two adaptation paths—textual and visual—to seamlessly integrate the two persona features, P^* and V^*.

The structure of the Personalized Text Encoder ($\mathcal{X}$) is shown in Fig. 3 (a). Given an input sentence like "A person is dancing" (T_{in}), we introduce a personalized sentence "[P*] person is dancing" ($\tilde{T}_{in}$), where the word [P*] precedes the subject. Subsequently, the vector P^*, derived from $\mathcal{E}$, replaces the token embedding corresponding to [P*]. Then, the modified sentence embedding, $\mathcal{X}_{clip}(\tilde{T}_{in})$, is adaptively merged with the original sentence embedding, $\mathcal{X}_{clip}(T_{in})$. The descriptions in HumanML3D are generally structured with subjects like 'A Person,' or 'Someone' and lack adjectives to characterize the subject. Thus, a smoother adaptation is necessary to enable the pretrained model to handle variations in sentence structure effectively. Specifically, personalized text feature T^* is derived as

$$\begin{aligned} T^* &= \mathcal{X}(T_{in}, P^*) \\ &= \mathcal{X}_{clip}(T_{in}) + s_t \cdot \tanh(\gamma_t) \cdot \mathcal{X}_{clip}(\tilde{T}_{in}, P^*), \end{aligned} \quad (4)$$

where γ_t is a learnable parameter initialized as zero which is for zero gating [36]. s_t is a scaling factor [31], a constant used to balance the adaptive layer during inference. CLIP text encoder $\mathcal{X}_{clip}$ is frozen during training.

For visual adaptation, we introduce a single adaptive layer into the transformer structure, following the approach used in [31] for the UNet. As illustrated in Fig. 3 (b), for b-th transformer block in t-th step of the diffusion, the adaptive layer is placed between the original self-attention layer and the feedforward layer as

$$z' = z + s_v \cdot \tanh(\gamma_v) \cdot \text{Adapt}([z, V^*]), \quad (5)$$

where Adapt indicates the adaptive layer and z and z' denote the input and output features of the adaptive layer, respectively. γ_v and s_v serve as the gating parameter and scaling factor. We evaluated several adaptive layer structures, including self-attention, cross-attention, and the AdaIN layer, which is widely used in style transfer [1, 12]. Among these, self-attention was found to be the most effective.

3.3. Training

We used a pretrained 50-step MDM [33] for $\mathcal{D}$. Diffusion is modeled as a Markov noising process $\{M^t\}_{t=0}^{T}$, where M^0 is sampled from the training motion, and t denotes the noise time-step. The diffusion loss for finetuning is given by

$$L_D := \mathbb{E}_{M^0,t,T} \left[\left\| M^0 - \mathcal{D}\left(M^t, t, V^*, T^*\right) \right\|_2^2 \right] + L_{geo}, \quad (6)$$

where L_{geo} is the geometric loss used during the MDM pretraining [33]. The final training loss is expressed as

$$L = L_D + \lambda L_{pc}, \quad (7)$$

where λ is a weight for the L_{pc}.

We utilize Classifier-Free Guidance (CFG) [8], which is a technique for adjusting the trade-off between the diversity and fidelity of generated samples. Similarly, we applied CFG during finetuning to enable control over the trade-off between the pretrained diffusion model's capability for generating diverse motions and fidelity to the newly provided conditions, V^* and T^*. During training, V^* and T^* are randomly dropped with a 10% probability, respectively. During sampling, the output of t-th step is extrapolated as

$$\begin{aligned} &\hat{\mathcal{D}}(M^t, t, V^*, T^*) = b\mathcal{D}_T + (1-b)\mathcal{D}_V, \quad (8) \\ &\mathcal{D}_T = \mathcal{D}(M^t, t, V^*, \emptyset) + g_t(\mathcal{D}(M^t, t, V^*, T^*) - \mathcal{D}(M^t, t, V^*, \emptyset)), \\ &\mathcal{D}_V = \mathcal{D}(M^t, t, \emptyset, T^*) + g_v(\mathcal{D}(M^t, t, V^*, T^*) - \mathcal{D}(M^t, t, \emptyset, T^*)), \end{aligned}$$

where $\emptyset$ indicates the absence of an input. g_t and g_v represent guidance scales, while b serves as the balancing factor between the modalities. As g_v increases, the model puts more focus on the persona inherent in the visual feature. However, if g_v is set too high, inconsistencies with the

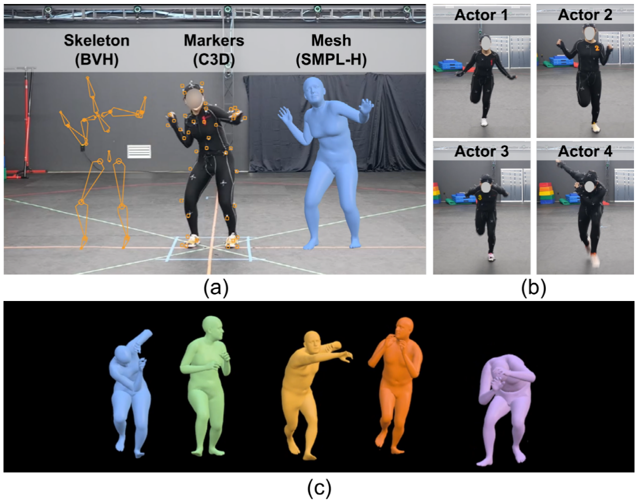


Figure 4. (a) Motion capture studio and examples of data formats: skeleton, markers, and mesh. (b) Unique persona expressions of each actor in the 'Childish' category. (c) Rendered mesh for each actor in the 'Fearful' category

text prompt may arise. On the other hand, g_t controls the model's adherence to the prompt, with the intensity of P^* in the text being adjusted by s_t in Eq. (4). The hyperparameter settings are specified in Sec. 5, and an ablation study for these is provided in the supplementary material.

3.4. Context-Aware Fusion for Multiple Inputs

Intuitively, if multiple motions are input, these additional cues can help with generating motion. However, it's also crucial to choose the right cues, as transferring persona from motions with very different content types, even with L_{pc}, can result in unnatural movements. A straightforward approach might be to take the mean of all features, as suggested in [31]. In the motion domain, however, taking all motions can result in blended motions that appear implausible. To overcome this, we introduce a Context-Aware Fusion (CAF) method that prioritizes input motions based on their contextual relevance to the input prompt. To assess this relevance, we use the motion encoder and text encoder of the motion clip ($\mathcal{E}clip$ and $\mathcal{X}clip$), which align the text and motion feature spaces. The CAF is formulated as

$$I_{\text{Top-k}} = \text{argmax}_i(S_i), \quad S_i = \text{sim}(\mathcal{E}_{\text{clip}}(M_i), \bar{T}^*), \tag{9}$$

$$w_i = \begin{cases} \dfrac{\exp(S_i)}{\sum_n \exp(S_n)}, & \text{if } i \in I_{\text{Top-k}}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\bar{T}^*$ represents the personalized text feature, using the average value of $\{P^*\}$ as a persona token. $\text{sim}(\cdot)$ denotes cosine similarity. Consequently,

$$V^* = \sum_i w_i V_i^*, \quad P^* = \sum_i w_i P_i^*. \tag{11}$$

Table 2. Comparison with existing motion style datasets. The note for '*' is provided in the main text. '-' indicates no information

Dataset	Actors	Styles	Contents	Clips	Mesh	Text
Xia [34]	-	8	6	572	✗	✗
BFA [1]	1	16	9	32	✗	✗
BN-1 [13]	2*	15	17	175	✗	✗
BN-2 [13]	2*	7	10	2,902	✗	✗
100Style [19]	1	100*	8*	810	✗	✗
PerMo(ours)	**5**	34	10	**6,610**	✓	✓

4. PerMo Dataset

We collected a large-scale PerMo dataset, capturing personas from multiple actors. To ensure variety, we hired five professional motion capture actors of diverse genders and body types. Each actor is assigned to perform 34 styles, categorized into *Age, Character, Condition, Emotion, Traits*, and *Surroundings*, resulting in a total of 170 personas. Each actor performed 10 distinct contents for every style, carefully selected to engage different body parts.

The dataset was captured in a studio equipped with 33 OptiTrack cameras, using 41 optical markers per person. For compatibility with other motion datasets, we converted the marker data to SMPL-H [29] format using Mosh++ [18]. All data underwent expert cleaning and verification, with the entire process adhering to strict quality guidelines. Fig. 4 (a) illustrates the motion capture studio and the provided data formats, while (b) highlights the distinct personas expressed by each actor. Additionally, we constructed 20-30 varied descriptions for each content type, following research indicating that high-quality finetuning descriptions improve prompt-based control [7]. Starting with one detailed description, we instructed ChatGPT to generate a range from high-level descriptions (e.g., "A person is jumping ahead") to low-level descriptions (e.g., "A person pushes off the ground and jumps forward several times"). Additional details on data categories, capture process, data structure, and examples can be found in the supplementary.

Table 2 shows a comparison with existing motion style datasets. Notably, PerMo is the first dataset to collect data from multiple actors. Our dataset also offers the highest number of total clips and content categories among trainable datasets and is the only one that includes mesh, marker data, and descriptions. The BN datasets [13] involve two actors; however, for the same category, only one actor is used (BN-1), or there are no tags identifying the actors (BN-2). BN-1 is unsuitable for training, as it contains only a single motion sequence per category. The 100Style [19] offers the largest variety of styles, but 58 of its categories focus on content, rather than on style [38]. Additionally, the eight content categories in this dataset mainly consist of walking or running, offering limited variation.

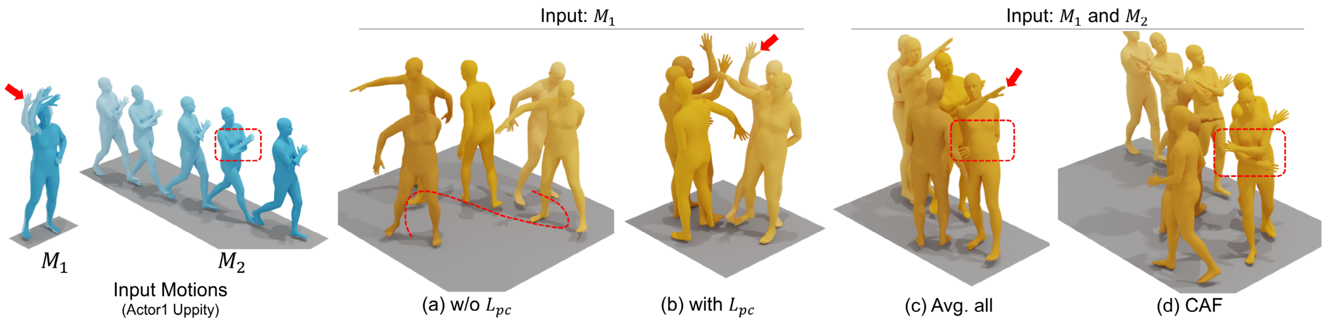


Figure 5. Example of the ablation study. The input motions are from the ‘Uppity’ of Actor 1. The input prompt is “A person walks in a circle.” In (a) and (b), only M_1 is provided for the input, while both M_1 and M_2 are provided for (c) and (d). L_{pc} encourages the generated motion to closely follow the prompt, while CAF prevents the motion from blending. We set $k = 1$ for CAF

5. Experiments

Implementation Details. We conduct experiments using the proposed PerMo and 100Style [19] datasets. During training, a single motion is cropped differently to serve as both the input and the GT motion for generation. We set λ to 10^{-2} and use the AdamW [16] optimizer with a learning rate of 10^{-4}. Training runs for 500 epochs with a batch size of 64. We set s_t and s_v to 1 during training and to 0.3 during inference. g_t is set to 10, and g_v to 15 and 10 for PerMo and 100Style, respectively. We set b to 0.7 for single-input and 0.5 for multi-input settings, and k to 5 for CAF.

Evaluation Settings. For quantitative evaluation, we use descriptions from the HumanML3D test set as prompts and motion samples from the PerMo and 100Style datasets as inputs. For each description, $[P^*]$ is logically placed as a modifier for the subject. We used two evaluation settings: Single Input (SI) and Multiple Input (MI). In SI, one motion is sampled from the entire dataset for comparison with MST methods.In MI, a persona set is first sampled, and then $|M_i|$ motions are drawn from it.

Performance is evaluated using Frechet Inception Distance (FID), R-Precision, and Diversity—metrics commonly used in motion generation tasks [3, 5, 6, 33]—as well as Persona Recognition Accuracy (PRA), a metric adopted by Style Recognition Accuracy (SRA) in MST [9, 32]. **FID** assesses the overall quality and realism of generated motions and is our primary metric [5]. **R-Precision** evaluates text-to-motion alignment accuracy [6], and **Diversity** [5] measures how well the generated motions reflect a broad distribution of personas. **PRA** uses a pretrained persona classifier to evaluate the persona consistency of generated motions. The details are in the supplementary material.

5.1. Ablation Study

Table 3 demonstrates the effects of our key contributions. The baseline represents results obtained using only visual adaptation. When we introduce P^* for textual adaptation, the FID score drops significantly, and both PRA and Diversity increase. This indicates that multimodal integration of persona information not only strengthens the reflection of persona but also supports the generation of plausible motion. Furthermore, adding L_{pc} leads to an additional reduction in FID and a notable increase in R-precision. This implies that L_{pc} helps the Persona Extractor effectively capture persona essence disentangled from the input content, resulting in generated motion that aligns well with the prompt without compromising content.

Table 3. Ablation study of the proposed components. $|M_i|$=max indicates that all motions in the chosen persona set are input

Methods	FID ↓	R Precision ↑			PRA avg.↑	Diversity ↑
		Top 1	Top 2	Top 3		
Single Input (SI) Setting						
baseline	7.45	0.06	0.11	0.16	17.99	7.48
+ P^*	5.06	0.05	0.10	0.15	**18.26**	**8.01**
+L_{pc}	**3.18**	**0.15**	**0.26**	**0.33**	18.05	7.74
Multiple Input (MI) Setting						
$\lvert M_i\rvert$=1	4.27	0.08	0.15	0.20	16.20	7.71
$\lvert M_i\rvert$=max, [31]	3.52	**0.19**	**0.30**	0.38	**19.24**	7.88
$\lvert M_i\rvert$=max, CAF	**2.95**	**0.19**	**0.30**	**0.39**	18.13	**8.12**

Next, the effectiveness of CAF is validated in the MI setting. When multiple input motions are used ($|M_i|$=max), all metrics improve compared to using a single input ($|M_i|$=1) due to the higher likelihood of finding motion candidates that align with the prompt. Additionally, even in cases with multiple inputs, selecting the contextually most relevant Top-k motions—as done by CAF—is shown to be crucial. For $|M_i|$=max, [31] presents a baseline approach that uses the average feature of all inputs. Introducing the proposed CAF reduces FID while increasing Diversity, indicating that the generated motions are sufficiently diverse and plausible rather than oversimplified by averaging. Although the PRA score decreases, we found through qualitative results that generating plausible motion is a more crucial factor in improving motion quality.

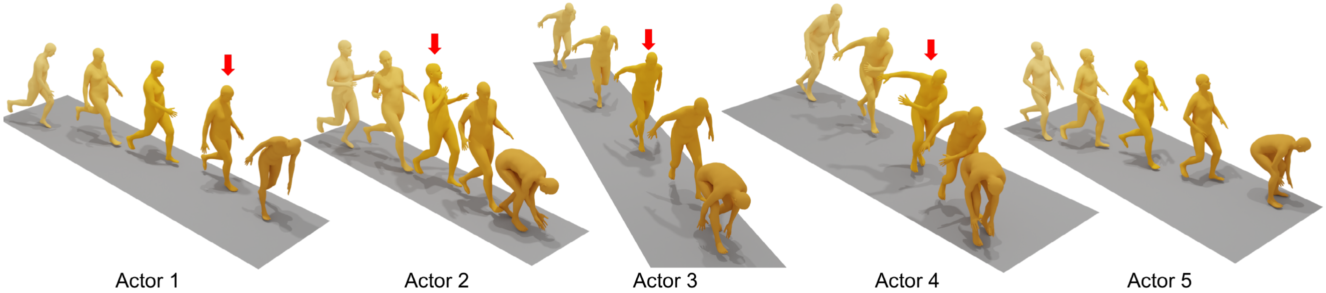


Figure 6. Example reflecting the different personas of Actors 1-5 for the same 'Childish' category. The input prompt is "A person is running and bending to pick something." Red arrows highlight the frames that clearly show distinct movements. The input motion of each actor can be referenced in Fig. 4 (b)

Table 4. Ablation study regarding adaptation. PersonaBooth (SI) indicates our complete model for single inputs

Methods	FID ↓	R Precision ↑			PRA avg.↑	Diversity ↑
		Top 1	Top 2	Top 3		
AdaIN	8.77	0.04	0.07	0.10	15.10	7.33
Cross-Attn	8.96	0.04	0.07	0.10	16.37	7.62
w/o T adapt.	3.61	0.14	0.23	0.31	17.77	**7.88**
PersonaBooth (SI)	**3.18**	**0.15**	**0.26**	**0.33**	**18.05**	7.74

Fig. 5 provides an example. When using only M_1 as the input motion, if L_{pc} is not applied, the generated motion fails to follow the path provided in the prompt, as shown in (a). However, with L_{pc}, the motion better aligns with the prompt's trajectory, though it still includes unintended hand-waving motions unrelated to the prompt. When using multiple input motions, including M_2 (a walking motion with arms crossed), taking the mean of features results in a blended motion, as shown in (c). This blending causes one arm to bend while the other hangs low, creating unnatural poses. In contrast, with CAF, M_2 receives a higher weight, resulting in a more natural walking motion with arms crossed, as shown in (d).

Table 4 presents various ablation results related to adaptation. For the visual adaptation layer, Self-Attention proved to be the most effective compared to AdaIN and Cross-Attn. Additionally, introducing text adaptation at Personalized Text Encoder improves the PRA, FID, and R-precision metrics. Fig. 6 demonstrates how PersonaBooth effectively captures and represents each actor's unique action. The red arrows specifically highlight the distinct arm movements of each persona. Additional details, including variations in body tilt and movement speed, are more clearly observable in the supplementary video materials.

5.2. Comparison with State-of-the-Art methods

We compare our PersonaBooth with the state-of-the-art diffusion-based MST approaches, MoMo [25] and MCM-LDM [32], on the PerMo dataset. Additionally, we evaluate it on an existing style dataset, 100Style [19]. As MCM-LDM does not support prompt-based generation, we used MDM-generated motions [33] as target motions for evaluation. MCM-LDM was evaluated using the original model trained on HumanML3D, as well as finetuned models, referred to as MCM-LDM, which were trained on the PerMo and 100Style datasets.

5.2.1. PerMo Dataset

The comparative results for the PerMo dataset are shown in Table 5, with all subcategory results displayed for PRA. The single-input PersonaBooth significantly reduces FID compared to existing methods, indicating that while previous studies struggled with PerMo—a challenging dataset containing more diverse content than 100Style—PersonaBooth performs robust in generating high-quality motion. Additionally, PersonaBooth achieves the highest performance in R-Precision, PRA, and Diversity compared to existing methods. Applying multiple inputs and CAF further enhances performance. It's noteworthy that while MoMo and MCM-LDM utilized a 100-step and 1000-step diffusion model, respectively, we used a 50-step diffusion model.

Fig. 7 shows qualitative results. MoMo [25] notes that their generated motions often follow the input motion's direction instead of the target's, which they attempted to copy the root rotation from the input motion (shown in (a) before and (b) after adjustment). The prompts for "turns in the air" were not met in (a), (b), or (c). Particularly, (c) does not reflect the style of ballerina. In contrast, PersonaBooth accurately generates the motion by first performing a hop that reflects the style, followed by a precise turn in the air. Example videos are provided in the supplementary material.

5.2.2. 100Style Dataset

Table 6 presents the comparative results on the 100Style dataset. MoMo demonstrates high PRA and Diversity, indicating effective style reflection. However, it shows very low R-precision, suggesting poor alignment between generated motions and prompts. Finetuned MCM-LDM improved performance on nearly all metrics compared to its

Table 5. Comparison with the state-of-the-art methods on the PerMo dataset. The comparison is made in the Single Input (SI) setting as the existing methods do not support multiple inputs. MCM-LDM* indicates the model is finetuned on the PerMo dataset.

Methods	Steps	FID ↓	R Precision ↑			PRA ↑										Diversity ↑
			Top 1	Top 2	Top 3	Age	Char1	Char2	Cond1	Cond2	Emo1	Emo2	Trait	Sur	**Avg.**	
MoMo [25]	100	13.91	0.05	0.09	0.13	28.23	10.23	11.30	**8.73**	15.97	9.97	14.57	12.60	7.93	13.47	6.74
MCM-LDM [32]	1000	9.16	0.13	0.23	0.30	47.38	11.44	10.34	7.63	18.90	11.69	17.20	15.71	**9.22**	17.00	6.70
MCM-LDM* [32]	1000	9.70	0.13	0.23	0.30	46.49	12.12	9.22	7.95	**19.77**	11.75	18.05	15.18	7.33	16.76	6.76
PersonaBooth (SI)	50	**3.18**	**0.15**	**0.26**	**0.33**	**48.00**	**13.67**	**11.97**	6.69	18.67	**14.34**	**19.75**	**17.06**	8.80	**18.05**	**7.74**
PersonaBooth (MI)	50	2.95	0.19	0.30	0.39	51.95	14.89	12.00	5.16	22.95	9.44	18.14	15.45	10.09	18.13	8.12

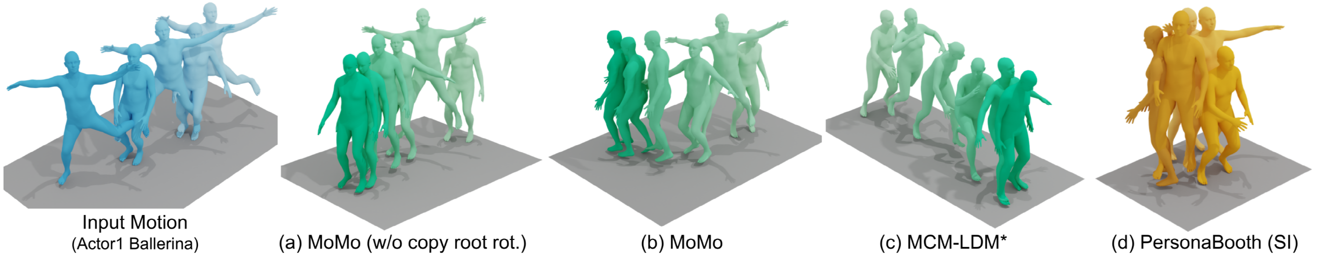


Figure 7. Qualitative comparison on PerMo dataset. The input prompt is “A person hops forward and turns in the air.”

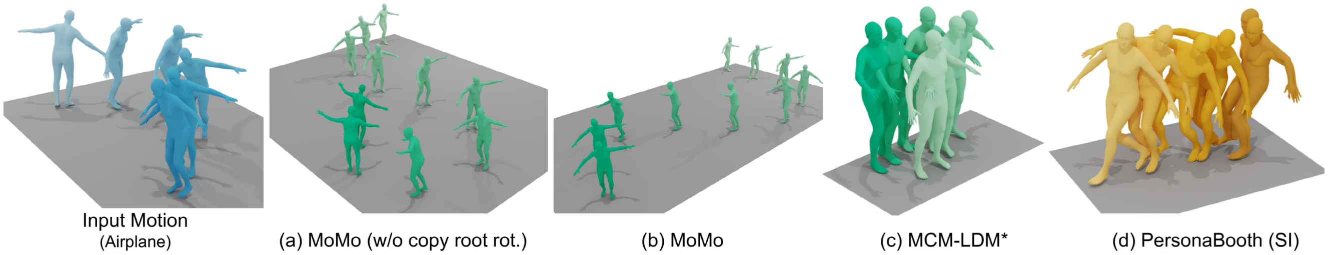


Figure 8. Qualitative comparison on 100Style dataset. The input prompt is “A person walks backward and sits down on the chair.”

Table 6. Comparison with the state-of-the-art methods on the 100Style dataset. MCM-LDM* indicates the model is finetuned on the 100Style dataset

Methods	FID ↓	R Precision ↑			SRA ↑	Diversity ↑
		Top 1	Top 2	Top 3		
MoMo [25]	5.97	0.07	0.10	0.14	60.57	7.82
MCM-LDM [32]	7.20	0.18	0.29	0.37	54.44	6.76
MCM-LDM* [32]	6.53	0.17	0.28	0.37	54.61	7.12
PersonaBooth (SI)	**3.27**	**0.20**	**0.31**	**0.40**	**64.53**	**7.90**

original model. Although finetuned MCM-LDM achieved high R-precision, it also had a high FID, resulting in unnatural motion generation, with low PRA and Diversity indicating weaker style reflection. In contrast, our single-input PersonaBooth significantly reduces FID and achieves top performance across all metrics, including R-precision, PRA, and Diversity.

In Fig. 8 (a) and (b), MoMo produces sliding motions that don’t align with the prompt. MCM-LDM shows sideward movements. In contrast, PersonaBooth accurately reflects both the input motion’s style, like arm extension and leaning, and the sitting motion in line with the prompt.

6. Conclusion

We propose the PerMo dataset and the PersonaBooth framework for Motion Personalization. Through contrastive learning, we effectively capture the essence of a persona. Furthermore, we enhance the integration of persona traits into diffusion models using a multi-modal finetuning method. We also introduce the CAF, which effectively handles multiple inputs, ensuring cohesive representation of personalized motion characteristics.

Limitations and Future Works. The PerMo contains relatively short motion sequences, which sometimes lead to stationary poses at the end of generated motions of the same length. However, this issue does not occur when prompted with longer actions, suggesting that automatic adjustment of motion length remains a task for future work. In addition, we plan to investigate approaches for applying CAF to each sequential action within a prompt individually. For instance, applying CAF separately to ‘run’ and ‘jump’ in the prompt “A person runs and then jumps.”

Acknowledgement.
This work was supported by MSIT, Korea, under [RS-2022-00164860] and [RS-2024-00418641] supervised by IITP.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64–1, 2020. 2, 4, 5

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[3] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 2, 6

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[5] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3D human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 6

[6] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 2, 3, 4, 6

[7] Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Kun Wan, Helge Rhodin, and Ratheesh Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. *arXiv preprint arXiv:2311.04315*, 2023. 5

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[9] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022. 2, 6

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3

[11] Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. Global-local motion transformer for unsupervised skeleton-based action learning. In *European Conference on Computer Vision*, pages 209–225. Springer, 2022. 2

[12] Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. Most: Motion style transformer between diverse action contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1714, 2024. 2, 4

[13] Makito Kobayashi, Chen-Chieh Liao, Keito Inoue, Sentaro Yojima, and Masafumi Takahashi. Motion capture dataset for practical use of ai-based motion editing and stylization. *arXiv preprint arXiv:2306.08861*, 2023. 5

[14] Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual Conference on Computer Graphics and Interactive Techniques*, pages 491–500, 2002. 1

[15] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[16] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[17] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[18] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 5

[19] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022. 2, 5, 6, 7

[20] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–17, 2021. 2

[21] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 1

[22] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023. 3

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

[24] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1

[25] Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit H Bermano, and Daniel Cohen-Or. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. *arXiv preprint arXiv:2406.06508*, 2024. 2, 7, 8

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[29] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 5

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[31] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 2, 4, 5, 6

[32] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2024. 2, 3, 6, 7, 8

[33] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4, 6, 7

[34] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34 (4):1–10, 2015. 2, 5

[35] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2

[36] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 4

[37] Xulu Zhang, Xiao-Yong Wei, Wengyu Zhang, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *arXiv preprint arXiv:2405.05538*, 2024. 2

[38] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. *arXiv preprint arXiv:2407.12783*, 2024. 2, 5