Learning Text-to-3D Motion Generation from 2D Motion Semantics

Zhicheng Shi, Tuo Feng, Wenguan Wang*, and Yi Yang

The State Key Lab of Brain-Machine Intelligence, Zhejiang University, China
https://github.com/nibglb/SWSL

Abstract. Data scarcity is a major constraint for data-driven Text-to-3D Motion generation. Prior attempts exploit abundant 2D motion data to address it, but typically treat 3D generation as a post-hoc stage of 2D prediction, resulting in camera-pose-dependent multi-stage pipelines and underutilization of rich 2D motion semantics (*e.g.*, temporal dynamics). To respond, we propose SWSL, a **S**emantic-aware **W**eakly **S**upervised **L**earning framework that leverages 2D-motion semantics to directly optimize end-to-end Text-to-3D Motion models. SWSL comprises three modules: Semantic Space Construction (SSC), which first constructs a shared semantic space aligning text, 2D-motion, and 3D-motion; Semantic-aware Pseudo-label Enhancement (SPE), which then uses weak supervision from 2D-motion embeddings to refine 3D pseudo-labels; and Semantic-level Feedback Optimization (SFO), which finally supplies embedding-level feedback for optimization. SWSL offers: ***(i)*** *reliance reduction* on text-3D motion data by utilizing text-2D motion pairs; ***(ii)*** *2D data constraint looseness* by extracting semantics from single-view, easy-accessible 2D motions; and ***(iii)*** *scalability* along mainstream Text-to-3D Motion advancement. On HumanML3D, SWSL yields consistent FID reductions under both a domain-specific protocol and a more general TMR-based protocol: -0.012 and -0.017 on simple baselines, along with -0.010 on a complicated one in a domain-specific setting, and -0.048, -0.038, and -0.056 under TMR setting - demonstrating its efficacy, robustness, and generalization beyond the domain-specific bias.

Keywords: Human Motion Generation · 2D Data for 3D Tasks · Weakly Supervised Learning

1 Introduction

Recent years have witnessed a growing interest in human motion generation, driven by diverse demands in robotics [4], game development [5], and virtual reality [6]. Within this domain, text-based 3D human motion generation (Text-to-3D Motion) stands out as a pivotal task due to the rich semantics embedded in natural language. Current research [2,7] on Text-to-3D Motion has proceeded along two directions, *i.e.*, designing sophisticated model architectures [7–23] and

* *Corresponding author.*

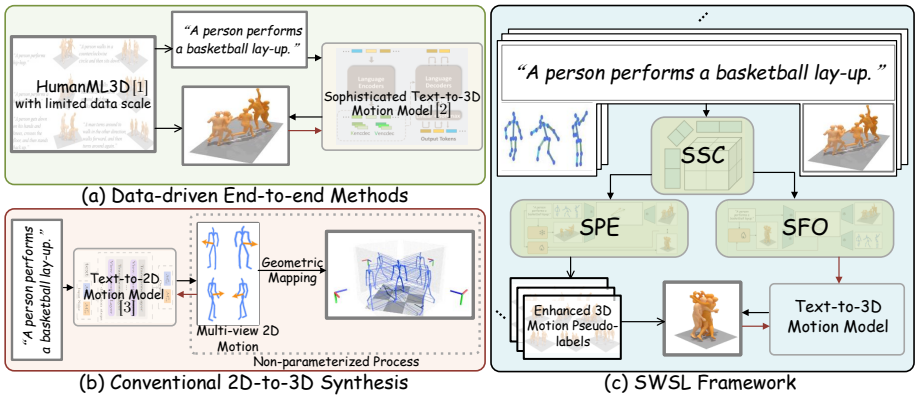

Fig. 1: Illustration of comparison (§1). (a) End-to-end 3D generation methods, *e.g.*, [2], rely heavily on text-3D motion datasets, *e.g.*, HumanML3D [1], which are limited in scale. (b) Existing 2D-based 3D generation methods, *e.g.*, [3], synthesize 3D motion via non-parameterized, geometric mapping process. (c) SWSL optimizes end-to-end methods [2] by leveraging 2D motion via three semantic-aware modules, namely SSC, SPE, and SFO (*cf*. §3.2). ⇢, →, and ⇢ denote data flow, parameter optimization, and condition acquisition, respectively.

scaling up model sizes [2, 24–28]. Despite technical strides, data-driven Text-to-3D Motion models heavily rely on vast amounts of paired text-3D motion data [1, 29] (*cf*. Fig. 1(a)). However, the collection of such data is prohibitively expensive due to the high costs of 3D motion-capture data acquisition and manual annotation. This leads to scarcity of text-3D motion data and potentially limits the progress of related research and applications [30].

Some pioneering endeavors [3, 30, 31] acknowledge this issue and attempt to mitigate it by applying easily accessible 2D motion data to the text-to-3D motion task (*cf*. Fig. 1(b)). However, these works still have the following limitations: ***(i)*** *Camera-pose-dependent multi-stage pipeline*: While these methods use 2D motion data in innovative ways, they commonly follow a pipeline that first predicts multi-view 2D motions and then synthesizes 3D motions using non-parameterized geometric techniques, *e.g.*, triangulation. This multi-stage, multi-view inference pipeline demands high-precision camera-pose parameters; ***(ii)*** *Overall semantics omission*: These approaches focus solely on the spatial geometric mapping between 2D and 3D motions, and the higher-level semantics encoded in 2D motions (*e.g.*, temporal dynamics) are underexplored, which limits the contribution of 2D motions to 3D motion generation.

To address these problems, we propose a 2D-motion-based semantic-aware weakly supervised learning framework, *i.e.*, SWSL (*cf*. Fig.1(c)) for Text-to-3D Motion. It consists of three sequentially-arranged modules: Semantic-aligned Space Construction (SSC), Semantic-aware Pseudo-label Enhancement (SPE) and Semantic-level Feedback Optimization (SFO). Specifically, in SSC, we first construct a semantic embedding space that aligns text, 2D motions, and 3D motions. Then, in SPE, SWSL maps the pseudo-labels from a primitively pre-

trained Text-to-3D Motion model, together with the 2D motion labels, into the embedding space and merges them into a new semantic vector, which is decoded into 3D motion; this yields the enhanced 3D pseudo-labels weakly supervised by 2D motion. Finally in SFO, with the same text-2D motion pairs, SWSL measures the discrepancy between 3D motion predictions and 2D motion labels at the embedding level; this discrepancy serves as a feedback signal to optimize the whole Text-to-3D Motion model. In summary, SWSL extracts semantic information of 2D motions as weak supervision and integrates it into the end-to-end training of Text-to-3D Motion models.

Our SWSL offers several distinctive features: ❶ It *reduces reliance on text-3D motion pairs* during Text-to-3D Motion training process, allowing 2D motion data to play a similar role to 3D motion data and thereby alleviating *text-3D motion data scarcity*. ❷ It is a *semantic-aware* training framework with *loosened 2D data constraint*. As SWSL constructs a shared semantic embedding space to capture motion dynamics and language semantics rather than keypoint geometry, it is less sensitive to camera viewpoint. This design mitigates *camera-pose-dependent multi-stage pipeline and overall semantics omission*. ❸ SWSL is a *general* framework that can be integrated seamlessly with the mainstream Text-to-3D Motion models, which allows SWSL to continuously adapt in step with advances in model architectures.

For comprehensive evaluation, we apply SWSL to representative Text-to-3D Motion baselines spanning both simple models (*i.e.*, T2M-GPT [32] and DiP [33]) and a complicated architecture (*i.e.*, Momask [34]). We assess performance under two protocols – a domain-specific setting and a more general TMR [35] protocol - on the gold-standard HumanML3D [1] benchmark. Under the domain-specific evaluation, SWSL yields consistent FID reductions of -0.012 and -0.017 on SB [32] and DiP* [33], along with -0.010 on Momask* [34]. Under TMR-based evaluation, the reductions are amplified to -0.048, -0.038, and -0.056, indicating generalization beyond domain-specific bias. Extensive experiments further show consistent improvements across diverse architectures. This not only reaffirms the utility of SWSL, but also highlights its practicality in real-world scenarios where 2D data sources are more readily available than 3D.

2 Related Work

Text-driven Human Motion Generation aims to generate natural, accurate and realistic 3D human motions from text descriptions. Recent text-to-motion advancement can be categorized into three mainstreams: (***i***) Cross-modal latent-alignment methods [35–38] employ VAEs [39] to align text and motion in a shared latent space and use a decoder to decode the latent features to motions. (***ii***) Diffusion-based methods [33, 40–48], *e.g.*, DiP [33], use Denoising Diffusion Process to predict motion distribution from Gaussian noise, with textual description as context. (***iii***) Vector-Quantization (VQ)-based methods [32, 34, 49–55], *e.g.*, T2M-GPT [32], leverage a discrete VAE [34, 56] to tokenize motions and employ

transformer-based networks to predict the next motion token conditioned on textual context and preceding motion tokens. Recent VQ-based methods [2, 24, 25] also utilize pre-trained large language models (LLMs) to generate motion tokens.

Though promising, prior works chiefly concentrate on sophisticated architecture designs to optimize generation quality (*cf*. Fig.1(a)). For example, Momask [34] adopts a hierarchical quantization scheme with two well-designed transformers, and HGM^3 [52] improves Momask by incorporating teacher–student Hard Token Mining and a text-based hierarchical semantic graph. However, these data-driven methods require large volumes of 3D motion data [1, 29], which are expensive to collect and annotate. The limited data availability inherently prevents these methods from achieving better performance [30], which our framework aims to address by leveraging 2D data to further optimize 3D models.

2D Data for 3D Generation. Applications of 2D motion data in 3D human motion generation mainly fall into three categories: ***(i)*** pose estimation, *e.g.*, SMPLify [57], which recovers 3D SMPL [58] pose parameters from 2D images. ***(ii)*** motion representation learning, *e.g.*, MotionBERT [59], which learns motion representations by recovering 3D motion from corrupted 2D skeleton sequences. ***(iii)*** motion generation, where methods such as MAS [30] exploit large-scale 2D motion corpora to train 3D motion generators and TENDER [31] expands this approach via a larger dataset. MVLift [60] learns a line-conditioned 2D motion diffusion prior and leverages multi-view geometric consistency to reconstruct 3D motion. Motion 2-to-3 [3] further demonstrates a text-conditioned pipeline that generates multi-view-consistent 2D motions and synthesizes them into 3D, illustrating a promising route for text-driven 3D motion generation.

Despite achieving promising results, [3, 30, 31] generate 2D motions end-to-end and then synthesize them into 3D motions via multi-view triangulation and reprojection, mainly focusing on spatial geometric consistency between 2D and 3D. In contrast, inspired by recent works in semantic mining [61, 62] and weakly supervised learning [63, 64], our method leverages 2D motion data at the semantic-embedding level to augment self-supervised pseudo-labels and mine feedback signals, directly optimizing end-to-end text-to-3D motion models and emphasizing overall semantics of 2D motion as inexact weak supervision [65].

Text-Motion Alignment. The success of image-text (*e.g.*, CLIP [66], BLIP [67] and COCA [68]), video-text (*e.g.*, Frozen [69] and MILNCE [70]) and audio-text (*e.g.*, MS2SL [71]) alignment models has prompted researchers to discover whether a cross-modal space exists between 3D human motions and language. T2M [1] uses a margin-based contrastive loss and batch-wise Euclidean distance to construct a retrieval model for evaluation. TMR [35] employs a VAE architecture to obtain a text-3D motion-aligned embedding space with an InfoNCE loss and negative filtering to attain promising alignment. Yu et al. [72] explore pretrained vision Transformer to align text and motion via an innovative patch-based motion representation. LaMP [73] trains a language–motion pretraining model via multi-task learning, which can be applied to text–motion alignment. In this work, we align 2D motions to text-3D motion-aligned space provided

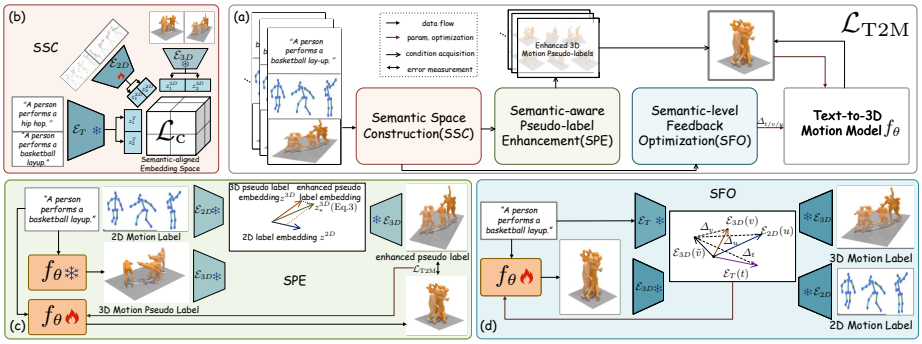

Fig. 2: Overview of 2D-motion based Semantic-aware Weak Supervision Learning (§3.2). As shown in (a), SWSL consists of three semantic-aware modules, SSC, SPE and SFO, which are displayed in (b), (c) and (d), respectively (Best viewed with zoom-in).

by TMR to capture 2D motion semantics, enabling direct semantic comparison between 3D and 2D at embedding level.

3 Method

As shown in Fig.2, we propose SWSL, a single-view 2D-motion-based Semantic-aware Weakly Supervised Learning framework, to leverage rich semantics from abundant single-view 2D motion data for Text-to-3D Motion model optimization. Firstly, §3.1 gives a brief preliminary of Text-to-3D Motion task. Secondly, §3.2 illustrates the general formulation of SWSL. SWSL is a principled training protocol without coupling to a specific model. Thirdly, §3.3 presents an instantiation of SWSL by adopting a representative architecture drawn from mainstream Text-to-3D Motion methods. Finally, implementation details are given in §3.4.

3.1 Preliminary

End-to-end Text-to-3D Motion. With a sentence describing a motion, Text-to-3D Motion is to generate a sequence of 3D human poses and trajectories that align with the textual context. More formally, given text-3D motion data distribution, denoted as $\mathcal{D}^{3D} = \{(t_n, v_n)\}_{n=1}^{N}$ with N samples, a Text-to-3D Motion model aims to convert the text input t_n to the corresponding human pose sequence v_n, *i.e.*, it aims to estimate the conditional distribution $f_\theta(v_n|t_n)$. With this estimation goal, the network parameters θ are optimized with Eq.1:

$$\theta^\star = \text{argmax}_\theta \sum\nolimits_{(t,v)\in\mathcal{D}^{3D}} \log P(v|t,\theta) = \text{argmin}_\theta \mathcal{L}_{\text{T2M}}(t), \tag{1}$$

where $\mathcal{L}_{\text{T2M}}(t)$ denotes target loss defined by the discrepancy between predicted 3D motion $\tilde{v}$ and ground truth v. As seen, the acquisition of Text-to-3D Motion models requires text-3D motion data pairs, which are limited in scale. Therefore,

in this work, we aim to further optimize Text-to-3D Motion mapping by utilizing text-2D motion pairs, which are easier to access than text-3D motion pairs [30].
Language Modality. Textual description T represents a written natural language sentence, which describes what and how a human motion is performed, *e.g.*, *"A person walks two steps and then stops"* or *"A person squats to the right"*.
Motion Modality. Human motion is defined as a sequences of human poses $H_{1:L} = (H_1^P, ..., H_L^P) \in \mathbb{R}^{L \times P}$. Each pose H_l^P corresponds to a kinematic chain of the articulated human body. For 3D motion, each sample is a sequence consisting of multiple temporally-ordered 3D human poses. For 2D motion, each sample is a sequence of single-view human poses.

3.2 2D-motion-based Semantic-aware Weakly Supervised Learning

We introduce Semantic-aware Weakly Supervised Learning (SWSL), a general training framework for text-to-motion models with abundant 2D motion data.
Semantic-aligned Space Construction (SSC). Existing 2D-based 3D generation pipelines [3,30,31] rely on non-parameterized synthesis grounded in spatial correspondences between multi-view 2D observations and 3D poses. In contrast, SWSL leverages the semantic information inherent in 2D motions by aligning text, single-view 2D motions, and 3D motions within a shared embedding space (*cf*. Fig 2(a)). We first construct a triplet of text, 2D motions, and 3D motions, and then contrastively train text/2D motion/3D motion encoders based on these triplets to obtain an aligned embedding space for all three. Given text-3D motion data $\mathcal{D}^{3D} = \{(t_n, v_n)\}_{n=1}^{N}$ and text-2D motion data $\mathcal{D}^{2D} = \{(t_m, u_m)\}_{m=1}^{M}$, we combine them into a text-2D motion-3D motion data $\mathcal{D}^{*} = \{(t_k, v_k, u_k)\}_{k=1}^{K}$ based on the semantic similarity of text modality between the datasets (*cf*. §3.3). Subsequently, three ACTOR-style [74] encoders $\mathcal{E}_T, \mathcal{E}_{3D}, \mathcal{E}_{2D}$ are defined for text, 3D motions, and 2D motions, respectively. Following [35], $\mathcal{E}_T$ encodes text using token embeddings from a pretrained language model, while $\mathcal{E}_{3D}$ and $\mathcal{E}_{2D}$ map the global trajectory and local kinematics into a compact embedding vector. $\mathcal{E}_T$ and $\mathcal{E}_{3D}$ are obtained from TMR [35] and $\mathcal{E}_{2D}$ is randomly initialized. Both $\mathcal{E}_T$ and $\mathcal{E}_{3D}$ are frozen and only $\mathcal{E}_{2D}$ is learnable. We train $\mathcal{E}_{2D}$ to learn 2D motion embedding by aligning them to well-learned 3D motion/language embeddings of TMR. Given a batch of B triplets $\{(t_b, v_b, u_b)\}_{b=1}^{B} \in \mathcal{D}^{*}$, the normalized latent vectors for sampled triplets are computed as $\{(z_b^{2D} = \mathcal{E}_{2D}(u_b)/|\mathcal{E}_{2D}(u_b)|, z_b^{3D} = \mathcal{E}_{3D}(v_b)/|\mathcal{E}_{3D}(v_b)|, z_b^{T} = \mathcal{E}_T(t_b)/|\mathcal{E}_T(t_b)|)\}_{b=1}^{B}$. With these latent vectors, the contrastive loss is then formulated as the following Eq.2:

$$\mathcal{L}_{\mathrm{c}} = -\sum_{b=1}^{B} \Big(\log \frac{\exp(z_b^{2D} \cdot z_b^{T}/\tau)}{\sum_j \exp(z_b^{2D} \cdot z_j^{T}/\tau)} + \log \frac{\exp(z_b^{T} \cdot z_b^{2D}/\tau)}{\sum_j \exp(z_b^{T} \cdot z_j^{2D}/\tau)} + \log \frac{\exp(z_b^{2D} \cdot z_b^{3D}/\tau)}{\sum_j \exp(z_b^{2D} \cdot z_j^{3D}/\tau)} + \log \frac{\exp(z_b^{3D} \cdot z_b^{2D}/\tau)}{\sum_j \exp(z_b^{3D} \cdot z_j^{2D}/\tau)} \Big), \tag{2}$$

where τ is a learnable temperature. The purpose of such triplet contrastive loss design is to construct an aligned embedding space, by pulling the positive triplet samples close and by pushing the negative triplet samples apart. After

this alignment process, a 2D motion $u \in \mathbb{R}^{L \times P_{2D}}$ can be mapped into the aligned space to obtain its semantic latent $z \in \mathbb{R}^W$, which are then used for Semantic-aware Pseudo-label Enhancement and Semantic-level Feedback Optimization.

Semantic-aware Pseudo-label Enhancement(SPE). With the text-2D-3D-aligned semantic space available, 2D and 3D motions can be mapped into the same embedding space for manipulation, *e.g.*, pseudo-label enhancement. Given a text-2D motion pair $(t_m, u_m) \in \mathcal{D}^{2D}$, a Text-to-3D Motion mapping f_θ pre-trained on $\mathcal{D}^{3D}$ is used to generate a pseudo 3D motion label $v_m = f_\theta(t_m)$. Subsequently, the 3D motion pseudo-label v_m, together with 2D motion label u_m are mapped into the semantic space, *i.e.*, $z^{3D} = \mathcal{E}_{3D}(v_m)$, $z^{2D} = \mathcal{E}_{2D}(u_m)$. These two semantic latents are merged into one via a weighted combination:

$$z_*^{3D} = z^{3D} + \lambda z^{2D}, \tag{3}$$

where $\lambda \in [0, 1]$ is an empirical weight, and z_*^{3D} is normalized regardless of λ. Finally, an ACTOR-style 3D motion decoder $\mathcal{G}_{\text{Dec}}$ is trained to recover z_*^{3D}, *i.e.*, $v_m^* = \mathcal{G}_{\text{Dec}}(z_*^{3D})$. The enhanced 3D motion pseudo-label v_m^* and its corresponding text t_m forms a new text-3D motion data distribution $\mathcal{D}^{3D*} = \{(t_m, v_m^*)\}_{m=1}^M$, which is added to $\mathcal{D}^{3D}$ to continue training 3D Motion model. Overall, the whole module integrates the semantics of 2D motion labels into the 3D motion pseudo-labels, yielding semantic-aware enhancement that helps improve the model's performance. Relevant qualitative results are later provided in §4.4.

Semantic-level Feedback Optimization (SFO). Since 2D and 3D motions are aligned in the same semantic space, the discrepancies in their corresponding semantic latents potentially reflect the semantic disparity between 2D and 3D motions, which may positively contribute to the optimization of Text-to-3D Motion network. Based on this motivation, we utilize the semantic latents of the mapping outputs to further optimize the Text-to-3D Motion model by minimizing discrepancies between the text, 3D motion, and 2D motion latents. More formally, given the triplet data $\mathcal{D}^* = \{(t_k, v_k, u_k)\}_{k=1}^K$, triplet encoders $\mathcal{E}_T$, $\mathcal{E}_{3D}$, $\mathcal{E}_{2D}$, and the Text-to-3D Motion mapping f_θ with enhanced pseudo labels, f_θ predicts the 3D motion $\tilde{v}_k = f_\theta(t_k)$. $\tilde{v}_k$ is expected to be aligned with t_k, v_k and u_k. Under this intra- and cross-modal consistency, the semantic-level errors $\Delta_t = \mathbf{d}(\mathcal{E}_T(t_k), \mathcal{E}_{3D}(\tilde{v}_k))$, $\Delta_v = \mathbf{d}(\mathcal{E}_{3D}(v_k), \mathcal{E}_{3D}(\tilde{v}_k))$, and $\Delta_u = \mathbf{d}(\mathcal{E}_{2D}(u_k), \mathcal{E}_{3D}(\tilde{v}_k))$ can be achieved, where $\mathbf{d}(\cdot, \cdot)$ is the distance measurement. These discrepancies are minimized for further model training.

3.3 Framework Instantiation

For clarity, we follow the notation in §3.2 throughout this section. The triplet data $\mathcal{D}^*$, the Text-to-3D Motion mapping f_θ, and the framework optimization are instantiated as follows:

Data Collection. Given a text-3D motion pair $(t_n, v_n) \in \mathcal{D}^{3D}$ and a text-2D motion pair $(t_m, u_m) \in \mathcal{D}^{2D}$, an external large language model $\mathcal{M}_{\text{LLM}}$ is used to compute cosine similarity score between t_n and t_m.

$$sim(t_n, t_m) = \frac{\mathcal{M}_{\text{LLM}}(t_n) \cdot \mathcal{M}_{\text{LLM}}(t_m)}{|\mathcal{M}_{\text{LLM}}(t_n)||\mathcal{M}_{\text{LLM}}(t_m)|}, \tag{4}$$

where $\mathcal{M}_{\text{LLM}}(t_n)$ and $\mathcal{M}_{\text{LLM}}(t_m)$ are sentence embeddings extracted by $\mathcal{M}_{\text{LLM}}$. Data pairs whose sentence-level cosine similarity exceeds an empirical threshold are merged into triplets to construct the triplet data $\mathcal{D}^*$.

Text-to-3D Motion Mapping. Latest methods [32,34] often follow a tokenization-prediction pipeline: 3D motions are first tokenized into discrete motion embeddings via a discrete VAE (*e.g.*, VQ-VAE [56] and RVQ-VAE [34]); Then, a seq2seq network (*e.g.*, autoregressive transformer) predicts tokenized motion, conditioned on text feature extracted by a large-scale pretrained text encoder (*e.g.*, CLIP [66]). These works adopt a representative architecture as follows:

A 3D motion tokenizer is first defined as a composition of an encoder $\mathcal{Q}_{\text{Enc}}$, a decoder $\mathcal{Q}_{\text{Gen}}$, and a learnt codebook Z. Then, a 3D motion sample v is fed into $\mathcal{Q}_{\text{Enc}}$ to obtain a latent feature sequence. Next, each element in this sequence is quantized by searching its nearest neighbor code vector in Z, yielding a corresponding sequence of indices $s = \text{Index}(Z, \mathcal{Q}_{\text{Enc}}(v))$, which represent the tokenized motion codes. $\text{Index}(\cdot,\cdot)$ is the nearest neighbor searching function. Finally, the original 3D motion is approximately reconstructed as $v' \approx \mathcal{Q}_{\text{Gen}}(Z[s])$, where $Z[s]$ maps each index in s back to its corresponding codebook vector.

After tokenization, an autoregressive next-index prediction network $\mathcal{F}_\theta$ is defined for Text-to-3D Motion. Given the previous $i-1$ indices $s_{<i}$, and the textual condition $\boldsymbol{c} = \mathcal{C}(t)$, which is extracted by the pretrained text encoder $\mathcal{C}$ from textual description t, $\mathcal{F}_\theta$ predicts the distribution of possible next index s_i: $\mathcal{F}_\theta(s_i|\boldsymbol{c}, s_{<i})$. The text t is the corresponding description for v. During training, only the parameters of $\mathcal{F}_\theta$ are trainable.

Framework Optimization. For SPE, we define the text-3D motion data $\mathcal{D}^{3D} = \{(t_n, v_n)\}_{n=1}^{N}$ along with the pseudo-labeled data $\mathcal{D}^{3D*} = \{(t_m, v_m^*)\}_{m=1}^{M}$, and train $\mathcal{F}_\theta$ on their union $\mathcal{D}^U = \mathcal{D}^{3D} \cup \mathcal{D}^{3D*}$. Given a batch of text-3D motion pairs $(t_b, v_b)_{b=1}^{B} \in \mathcal{D}^U$, $\mathcal{F}_\theta$ generates motion token sequences $\tilde{s}^{(B,L)}$ guided by $t_{b=1}^{B}$, *i.e.*, $\tilde{s}^{(B,L)} = \mathcal{F}_\theta(t_{b=1}^{B})$. The cross-entropy between generated motion tokens $\tilde{s}^{(B,L)}$ and quantized ground-truth motions $s^{(B,L)} = \mathcal{Q}_{\text{Enc}}(v_{b=1}^{B})$ is formulated as Eq.5 and is minimized to optimize $\mathcal{F}_\theta$:

$$\mathcal{L}_{\text{CE}}(s^{(B,L)}, \tilde{s}^{(B,L)}) = -\frac{1}{B}\sum\nolimits_{b=1}^{B}\sum\nolimits_{l=1}^{L} s_{(b,l)} \log \tilde{s}_{(b,l)}. \tag{5}$$

For SFO, $\mathbf{d}(\cdot,\cdot)$ is instantiated as batch-wise Frechet Inception Distance (FID) [75] between generated motion embeddings and ground truth. Given a batch of triplet data $\{(t_b, v_b, u_b)\}_{b=1}^{B} \in \mathcal{D}^*$, the generated motion tokens $\tilde{s}_{b=1}^{B} = \mathcal{F}_\theta(t_{b=1}^{B})$ are firstly decoded by $\mathcal{Q}_{\text{Gen}}$ to obtain motions $\tilde{v}_{b=1}^{B}$, and then encoded by $\mathcal{E}_{3D}$ to get semantic motion embeddings for discrepancy measurement $\Delta_{t/v/u}$. As the dequantization process truncates gradient flow, $\Delta_{t/v/u}$ are non-differentiable. Therefore, we further optimize $\mathcal{F}_\theta$ via RL and define an FID-based reward measuring semantic similarity between prediction and ground-truth:

$$\begin{aligned}\mathcal{R}(\tilde{v}_{b=1}^{B}) = -\,(&\omega_t \text{FID}(\mathcal{E}_T(t_{b=1}^{B}), \mathcal{E}_{3D}(\tilde{v}_{b=1}^{B})) \\ &+ \omega_v \text{FID}(\mathcal{E}_{3D}(v_{b=1}^{B}), \mathcal{E}_{3D}(\tilde{v}_{b=1}^{B})) \\ &+ \omega_u \text{FID}(\mathcal{E}_{2D}(u_{b=1}^{B}), \mathcal{E}_{3D}(\tilde{v}_{b=1}^{B}))),\end{aligned} \tag{6}$$

where $\omega_{t/v/u} \in [0, 1]$ are empirical weights. This reward is maximized via the following policy gradient:

$$\mathcal{L}_{RL} = -\mathbb{E}_{\tilde{v}_{b=1}^{B} \sim \mathcal{F}_\theta}[\mathcal{R}(\tilde{v}_{b=1}^{B})]. \tag{7}$$

The policy gradient descent uses SCST [76] with PPO [77]. The whole design enhances training stability and explores the semantics of single-view 2D motion with text-2D motion data throughout the whole optimization process.

3.4 Implementation Details

Network Architecture. We adopt VQ-VAE [56] and RVQ-VAE [34] as discrete VAE, CLIP [66] for text condition extraction, and employs [32] and [34] for $\mathcal{Q}_{\text{Enc}}$, $\mathcal{Q}_{\text{Gen}}$, Z and $\mathcal{C}$. Text-to-3D Motion network is implemented as a standard Transformer [78]. Text, 2D motion, and 3D motion encoders are Transformer encoders [78] with additional learnable distribution parameters, as in the VAE-based ACTOR [74]. An encoder outputs parameters of a Gaussian distribution (μ and σ), from which a latent vector $z \in \mathbb{R}^w$ can be sampled. The text encoder takes text features from a pretrained and frozen DistilBERT [79] as input, and the motion sequence is fed directly into the motion encoder. The 3D motion decoder is a Transformer decoder, same as that in ACTOR. It is learned to reconstruct the original 3D motion from its motion latent. $\mathcal{M}_{\text{LLM}}$ is implemented as a MPNet [80] network with pretrained weights.
Training. The empirical sentence similarity threshold for triplet data collection is 0.9. Triplet contrastive learning uses a batch size of 128. The empirical weight λ is set to 1.0. We pretrain the discrete motion VAE following the same training process to [32] and [34] and keep it fixed during training. The batch size for training $\mathcal{F}_\theta$ is 128 and the learning rate is 1×10^{-4}.
Reinforcement Learning. RL setup follows [81]. We treat the Text-to-3D Motion model as a policy model, and apply a KL penalty weighted by β to limit deviation from a frozen Text-to-3D Motion model trained via SPE. All Dropout modules are disabled during policy training [81,82]. The batch size, learning rate, β, γ, PPO epochs per batch is set to 128, 1×10^{-7}, 0.02, 1, and 4, respectively. $\omega_{t/v/u}$ is set to 0.2, 1.0, 0.2, respectively.
Inference. Following conventions [32, 34], we adopt probabilistic sampling to generate non-deterministic outputs for Text-to-3D Motion. Notably, we adopt the same inference settings as [32, 34] without adding extra parameters.
Reproducibility. All the experiments are conducted on four NVIDIA GeForce RTX 3090 GPUs with Pytorch 2.0.0.

4 Experiments

4.1 Experimental Setup

Benchmark. We conduct our experiments on HumanML3D [1], which is the gold-standard text-motion benchmark for text-conditioned 3D human motion

generation. It provides 14,616 unique human motions and 44,970 texts composed by 5,371 distinct words. Each motion sample is manually annotated with at least 3 descriptions and converted to a local rotation-based representation. All the motion clips are scaled to 20 frames per second.

2D Motion Data. 2D motion data $\mathcal{D}^{2D}$ is established with MotionVid [83] dataset. It comprises 1.2M single-view 2D motion videos of varying resolutions and durations exceeding 64 frames. The 2D human skeletons are extracted by DWPose [84]. We access the Motion-X [85] subset of MotionVid and randomly select 12K samples as $\mathcal{D}^{2D}$ to balance performance and training cost. Each human pose is represented in the OpenPose [86] format, including body, finger, and facial keypoints. To focus on semantic-rich torso movement, only body keypoints are retained. A pre-processing step similar to that of HumanML3D is applied to transform the 2D skeleton into a rotation-based representation.

Triplet Data Accession. Motion-X [85] and MotionVid [83] are employed to construct the triplet dataset $\mathcal{D}^*$. Motion-X comprises 81.1K diverse 3D motion clips (including locomotion, object interaction, dancing, and martial arts) paired with 81.1K semi-automatically annotated sequence-level text descriptions. We merge Motion-X and MotionVid through the process described in §3.3, yielding 15K triplets as $\mathcal{D}^*$. 3D motion samples from Motion-X follow the SMPL-X format [87], which includes body, finger, and facial keypoints. To align with the representation of HumanML3D, only the body keypoints are considered here.

Baseline Setting. For cross-modal alignment evaluation, we adopt standard text-3D motion alignment methods as baselines, and compare their text-3D alignment performance with SSC's text-2D and 3D-2D alignment results, thereby assessing the semantic alignment quality of the SSC embedding space. For text-conditioned 3D motion generation, we set three representative baselines, SB [32], Momask* [34] and DiP* [33], as follows. ***(i)*** A simple baseline [32], termed SB, is set according to §3.3. It utilizes VQ-VAE to tokenize motion and employs a Transformer decoder for next-token prediction. ***(ii)*** Following the same autoregressive pipeline, a more complicated baseline [34] uses a Residual VQ-VAE to tokenize motion into multi-layer tokens and adapts a bidirectional masked transformer to generate base-layer motion tokens via masking and remasking. This baseline is denoted as Momask*. ***(iii)*** We also set a simple diffusion-based baseline [33], denoted as DiP*, to comprehensively validate the efficacy of SWSL. This baseline generates motion from Gaussian noise using a transformer-based DDPM [88], with text feature and diffusion step serving as contextual input. Detailed framework instantiation of DiP* can be found in *Appendix* C.

Evaluation Protocol. Standard protocol evaluates model outputs in a text-3D motion-aligned embedding space for fidelity, diversity, and cross-modal alignment. Conventional work [1] constructs GRU-based text/motion encoders [89] and trains them from scratch on a specific benchmark with a margin-based contrastive loss. This introduces a scaling discrepancy and limits the generalizability of the evaluation. To address this, we also evaluate model outputs on TMR [35] setup that establishes a more generalized and normalized embedding space:

– **CONVENTION:** it uses dataset-specific, GRU-based encoders per benchmark with margin-based contrastive loss.
– **TMR:** it creates a more general and normalized embedding space by: ***(i)*** employing a fixed, general-purpose DistilBERT [79] to extract text features as the input of the text encoder; ***(ii)*** employing a VAE-based Transformer encoders as the text/motion encoders and an InfoNCE-style contrastive loss to mitigate the influence of scale.

Metrics. For cross-modal alignment, we follow [35] and report standard retrieval performance measures: R-Precision [90] and Median Rank (MedR) across the whole test set. For text-conditioned 3D motion generation, we follow conventions [32, 34, 90] and outline the metrics as three parts to pursue comprehensive evaluation: ***(i)*** Motion fidelity: Frechet Inception Distance (FID) [75], the primary metric, measures the distance of feature distributions between generated motions and ground truth. ***(ii)*** Motion diversity: we use Diversity and Multi-Modality (MModality) to reckon the diversity of generated motions. The former calculates the embedding-level output variance across the entire testing results, while the latter assesses the diversity of generated motions under identical texts. ***(iii)*** Text-motion alignment: the motion-retrieval precision (R-Precision) [90] determines the accuracy of text-motion matching utilizing top-1/2/3 retrieval accuracy. Multi-modal Distance (MM Dist) directly computes embedding-level distance between textual inputs and generated motions. Note that all these metrics derive from the text/motion-aligned space, and their values vary depending on the evaluation protocol.

4.2 Performance of SSC on Cross-modal Alignment

We conduct cross-modal retrieval experiments in the text-2D-3D semantic-aligned space to validate the effectiveness of SSC. A triplet version of HumanML3D test set, which contains texts, 3D motions, and their corresponding 2D motions, is established via the data collection process in §3.3 for a fair evaluation. The quantitative results are presented in Table 1. As shown, the retrieval accuracy of SSC in text-to-2D motion and 3D motion-to-2D motion is comparable to the text-to-3D motion retrieval performance reported by TMR [35]. This indicates that 2D motion aligns well with text and 3D motion in SSC, providing a reliable semantically aligned embedding space for subsequent SPE and SFO.

Method	Modality 1	Modality 2	1-to-2 Retrieval				2-to-1 Retrieval			
			R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
TEMOS [36] [ECCV22]	Text	3D Motion	2.12	8.26	13.52	173.0	3.86	9.38	14.00	183.25
T2M [1] [CVPR22]	Text	3D Motion	1.80	7.12	12.47	81.00	2.92	8.36	12.95	81.50
TMR [35] [ICCV23]	Text	3D Motion	5.68	20.34	30.94	28.00	9.95	23.56	32.69	28.50
Yu et al. [72] [CVPR24]	Text	3D Motion	8.46	23.56	35.27	23.00	9.63	22.87	33.57	25.00
LaMP* [73] [ICLR25]	Text	3D Motion	5.70	20.51	31.06	27.25	9.97	23.68	32.94	27.00
SSC (§3.2)	Text	2D Motion	5.22	18.47	29.13	26.50	9.03	21.88	30.56	27.25
	3D Motion	2D Motion	8.90	23.01	32.31	23.00	11.34	25.74	34.41	25.50

Table 1: Quantitative results on HumanML3D test set [1] in cross-modal alignment (§4.2). ∗ denotes our reproduced baselines. Blue denotes remarkable performance. Other results are directly borrowed from their respective papers.

Method	R-Precision ↑			FID ↓	MM Dist ↓	Diversity ↑	MModality ↑
	Top-1	Top-2	Top-3				
Real motions	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
TM2T [90] [ECCV22]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$	$2.219^{\pm.074}$
MDM [40] [ICLR23]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.779^{\pm.072}$
MLD [41] [CVPR22]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
T2M-GPT [32] [CVPR23]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
AttT2M [49] [ICCV23]	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$0.112^{\pm.006}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
MotionGPT [2] [NeurIPS23]	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$9.528^{\pm.071}$	$2.008^{\pm.084}$
AvatarGPT [24] [CVPR24]	$0.510^{\pm.005}$	$0.702^{\pm.005}$	$0.796^{\pm.003}$	$0.168^{\pm.008}$	-	$9.624^{\pm.055}$	-
EnergyMoGen [48] [CVPR25]	$0.526^{\pm.003}$	$0.718^{\pm.003}$	$0.815^{\pm.002}$	$0.176^{\pm.006}$	$2.931^{\pm.007}$	$9.500^{\pm.091}$	$2.270^{\pm.057}$
Motion 2-to-3 [3] [ICCV25]	-	-	$0.697^{\pm.000}$	$0.321^{\pm.000}$	$3.579^{\pm.000}$	$9.286^{\pm.000}$	-
SB [32]	$0.463^{\pm.004}$	$0.627^{\pm.002}$	$0.742^{\pm.005}$	$0.109^{\pm.002}$	$3.361^{\pm.010}$	$9.453^{\pm.034}$	$1.890^{\pm.048}$
SB + **SWSL**	$0.465^{\pm.003}$	$0.631^{\pm.002}$	$0.748^{\pm.004}$	$0.097^{\pm.004}$	$3.293^{\pm.008}$	$9.465^{\pm.042}$	$2.092^{\pm.037}$
Momask* [34]	$0.510^{\pm.004}$	$0.704^{\pm.005}$	$0.799^{\pm.004}$	$0.089^{\pm.002}$	$3.007^{\pm.009}$	$9.505^{\pm.045}$	$1.303^{\pm.040}$
Momask* + **SWSL**	$0.512^{\pm.003}$	$0.708^{\pm.004}$	$0.804^{\pm.003}$	$0.079^{\pm.003}$	$2.988^{\pm.008}$	$9.530^{\pm.037}$	$1.465^{\pm.029}$
DiP* [33]	$0.464^{\pm.003}$	$0.668^{\pm.002}$	$0.777^{\pm.004}$	$0.283^{\pm.000}$	$3.150^{\pm.008}$	$9.210^{\pm.030}$	$2.110^{\pm.046}$
DiP* + **SWSL**	$0.466^{\pm.002}$	$0.671^{\pm.003}$	$0.783^{\pm.003}$	$0.266^{\pm.003}$	$3.087^{\pm.007}$	$9.294^{\pm.034}$	$2.209^{\pm.053}$

Table 2: Quantitative Text-to-3D Motion comparison on HumanML3D test set [1] under CONVENTION protocol (§4.3). Protocol details are in §4.1. Metrics are computed using encoders from [1], which are biased to specific HumanML3D data distribution. ± indicates 95% confidence interval. Blue denotes SWSL improvements. (*) denotes our re-trained baselines (§4.1); other results are directly borrowed from respective papers.

Method	R-Precision ↑			FID ↓	MM Dist ↓	Diversity ↑	MModality ↑
	Top-1	Top-2	Top-3				
Real motions	$0.433^{\pm.002}$	$0.598^{\pm.003}$	$0.682^{\pm.002}$	$0.001^{\pm.000}$	$4.973^{\pm.010}$	$10.326^{\pm.023}$	-
SB [32]	$0.430^{\pm.002}$	$0.589^{\pm.003}$	$0.669^{\pm.003}$	$0.625^{\pm.002}$	$5.143^{\pm.011}$	$10.283^{\pm.055}$	$3.219^{\pm.035}$
SB + **SWSL**	$0.434^{\pm.003}$	$0.593^{\pm.002}$	$0.675^{\pm.004}$	$0.577^{\pm.003}$	$5.130^{\pm.016}$	$10.388^{\pm.033}$	$3.632^{\pm.029}$
Momask* [34]	$0.435^{\pm.002}$	$0.607^{\pm.004}$	$0.698^{\pm.004}$	$0.595^{\pm.002}$	$5.106^{\pm.010}$	$10.350^{\pm.045}$	$1.962^{\pm.034}$
Momask* + **SWSL**	$0.440^{\pm.003}$	$0.615^{\pm.004}$	$0.707^{\pm.005}$	$0.539^{\pm.002}$	$5.090^{\pm.029}$	$10.369^{\pm.035}$	$1.976^{\pm.035}$
DiP* [33]	$0.432^{\pm.003}$	$0.576^{\pm.002}$	$0.678^{\pm.003}$	$0.736^{\pm.004}$	$5.234^{\pm.026}$	$9.964^{\pm.027}$	$3.572^{\pm.051}$
DiP* + **SWSL**	$0.436^{\pm.002}$	$0.585^{\pm.002}$	$0.690^{\pm.003}$	$0.698^{\pm.003}$	$5.131^{\pm.026}$	$10.026^{\pm.021}$	$3.704^{\pm.056}$

Table 3: Quantitative Text-to-3D Motion comparison on HumanML3D test set [1] under TMR protocol (§4.3). Protocol details are in §4.1. Metrics computed with re-trained encoders on generalized and normalized embedding space.

4.3 Performance on Text-to-3D Motion

Quantitative Results. We summarize quantitative results on Text-to-3D Motion comparison under different evaluation protocols in Tables 2 and 3, respectively, from which we make three major observations: ***(i)*** Though affected by the biased CONVENTION protocol, models enhanced by SWSL achieve moderate improvements in FID. This result demonstrates that SWSL improves both simple and complex baselines in terms of generation fidelity, indicating that its performance gains scale with architectural improvements. Although three SWSL-enhanced models do not surpass the latest Text-to-3D Motion methods, *e.g.*, EnergyMoGen [48] and AvatarGPT [24] on R-Precision, MM Dist and MModality, they achieve comparable performance with it on the primary metric FID. Besides, improvements over the SOTA-FID method, *i.e.*, Momask* [34], show SWSL's potential to further improve them. ***(ii)*** Under the TMR protocol, we gain remarkable performance increases in R-Precision and MM Dist, further validating SWSL's effectiveness in improving Text-to-3D Motion models' ability to align text and motion. ***(iii)*** Under both protocols, SWSL boosts overall and

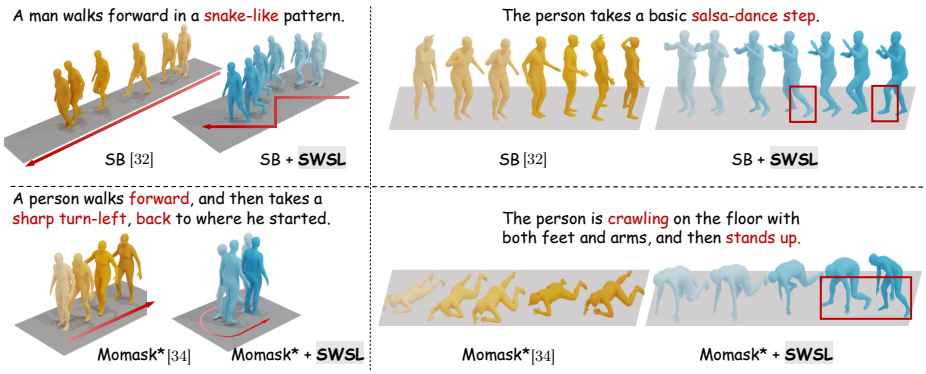

Fig. 3: Qualitative comparison between baselines and SWSL-improved models on Text-to-3D Motion generation (§4.3).

class-wise diversity of simple and complicated baselines. This shows that SWSL enhances models' capability to generate diverse and versatile 3D motions.

Qualitative Results. As displayed in the left part of Fig. 3, SB + SWSL successfully generates a walking motion with a snake-like, zig-zag trajectory, whereas SB [32] gives an ordinary straight-line locomotion. Besides, Momask* [34] ignores the description *"forward"* and *"turn-left"* and only generates a backward walking motion, while Momask* + SWSL outputs the fully-aligned motion with the forward-turn left-backward movement. These results indicate that SWSL improves the models' performance in fine-grained semantic perception and detailed motion generation. In the right part, the motion generated by SB mainly focuses on arm movements. In contrast, SB + SWSL produces a dancing motion with massive hip, waist, and gluteal movements, which are exactly the characteristics of Salsa dance. This shows that SWSL enhances the models' ability to generate a wider range of motions. Similarly, Momask* [34] produces a misaligned motion that shows a person leaning on the floor, while Momask* + SWSL predicts the correct motion with the crawling-and-standing-up movements. This demonstrates that SWSL boosts the models' overall semantic understanding.

4.4 Diagnostic Experiments

We perform a set of ablative studies with SB [32] under TMR protocol.

Training Strategy. Firstly, we validate the significance of both SPE and SFO in our SWSL framework. The first row of Table 4a reports the result of a bare baseline trained on benchmark data for comparison. In the second row, the SPE module is employed to train a Text-to-3D Motion model using additional pseudo-labels with 2D enhancement. This leads to notable improvement (-0.039 on FID), indicating that using additional volumes of pseudo labels with enhanced semantics can already improve the model's performance. In the third row, the SFO module without pseudo labels is used to introduce semantic-embedding-level text/2D/3D weak supervision, and this also gains observable improvements

(a) Training Strategy

Method	FID ↓	R-TOP1 ↑	R-TOP3 ↑	Diversity ↑	MModality ↑
SB [32]	0.625	0.430	0.669	10.283	3.219
SWSL *w.* SPE	0.586	0.432	0.673	10.334	3.486
SWSL *w.* SFO	0.590	0.431	0.671	10.325	3.353
Full SWSL	0.577	0.434	0.675	10.388	3.632

(b) 2D Data Size

Data Size	FID ↓	R-TOP1 ↑	R-TOP3 ↑	Diversity ↑	MModality ↑
3k	0.613	0.430	0.670	10.297	3.341
5k	0.598	0.428	0.671	10.309	3.450
9k	0.577	0.434	0.675	10.388	3.632
12k	0.574	0.436	0.677	10.407	3.665

(c) 3D Data Reliance

Data Ratio	FID ↓	R-TOP1 ↑	R-TOP3 ↑	Diversity ↑	MModality ↑
$\mathcal{D}^{3D}$	0.625	0.430	0.669	10.283	3.219
$1/2\mathcal{D}^{3D}+\mathcal{D}^{2D}$	0.630	0.426	0.664	10.241	3.201
$1/4\mathcal{D}^{3D}+\mathcal{D}^{2D}$	0.632	0.422	0.656	10.215	3.193
$1/8\mathcal{D}^{3D}+\mathcal{D}^{2D}$	0.691	0.405	0.625	10.014	3.007

(d) 2D Enhancement Efficacy

λ	FID ↓	R TOP1 ↑	R TOP3 ↑	Diversity ↑	MModality ↑
0.0	0.610	0.431	0.670	10.301	3.283
0.2	0.598	0.431	0.671	10.314	3.397
0.5	0.590	0.432	0.673	10.345	3.451
1.0	0.577	0.434	0.675	10.388	3.632

Table 4: Ablation studies on HumanML3D test set [1] under TMR (§4.1) evaluation protocol (§4.4).

(-0.035 on FID). This indicates that semantic-level discrepancy provides positive feedback for model optimization. As shown in the fourth row, SPE together with SFO is applied, and the biggest improvement (-0.048 on FID) is achieved. This demonstrates the efficacy of both modules in SWSL.

2D Data Size. Subsequently, we investigate the effectiveness of the 2D motion data size. Different volumes of 2D motion data are applied to SWSL for SB training. The results in Table 4b demonstrate that increasing the amount of 2D motion data consistently boosts performance. This evidences the scalability of SWSL with 2D motion data. To balance model performance gain and computational resource usage, 9K motion samples are chosen in SWSL by default.

3D Data Reliance. Furthermore, we consider the reliance of SWSL on 3D motion data. Different SWSL setups are employed, with a fixed amount of 2D motion data and a gradually decreasing amount of 3D motion data. As displayed in Table 4c, a 75% reduction in the volume of the 3D motion data incurs only a marginal -0.007 FID degradation. This indicates that SWSL indeed reduces reliance on 3D motion data.

2D Enhancement Efficacy. Finally, we present the efficacy of semantic-aware 2D enhancement on self-generated 3D pseudo labels in Table 4d. As seen, increasing λ from 0 to 1 consistently improves Text-to-3D Motion models. This indicates that merging 2D and 3D motion semantic vectors in the embedding space enhances pseudo-labels and reaffirms the efficacy of the SPE module.

5 Conclusion and Discussion

In this work, we propose a 2D-motion-based Semantic-aware Weakly Supervised Learning (SWSL) framework for text-based human motion generation. By fostering overall semantics of 2D motion data, SWSL reduces 3D data dependency, loosens 2D data constraints, enhances model performance, and shows promise for real-world applications where 2D data are more available than 3D data. SWSL is also a general training framework that shows scalability with architecture improvements. Performance gains, which are achieved by SWSL across three baselines under two evaluation protocols, demonstrate its efficacy. In future

work, we will extend SWSL to a broader range of Text-to-3D Motion baselines and datasets to ensure generalization, and incorporate additional 2D motion samples and computational resources to further improve both the conventional and the SOTA models.

Acknowledgements. This work was supported by Zhejiang Provincial Natural Science Foundation of China (No. LD25F020001), and Ant Group through CCF-Ant Research Fund.

References

1. 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 *CVPR*, pages 5152–5161, 2022.
2. Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, volume 36, 2024.
3. Ruoxi Guo, Huaijin Pi, Zehong Shen, Qing Shuai, Zechen Hu, Zhumei Wang, Yajiao Dong, Ruizhen Hu, Taku Komura, Sida Peng, and Xiaowei Zhou. Motion-2-to-3: Leveraging 2d motion data for 3d motion generations. In *ICCV*, pages 14305–14316, October 2025.
4. Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *RAS*, 109:13–26, 2018.
5. Jongmin Kim, Hoill Jung, MyungA Kang, and Kyungyong Chung. 3d human-gesture interface for fighting games using motion recognition sensor. *Wirel. Pers. Commun.*, 89:927–940, 2016.
6. Gregory F Welch. History: The use of the kalman filter for human motion tracking in virtual reality. *Presence*, 18(1):72–91, 2009.
7. Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *CVPR*, pages 27859–27871, 2025.
8. Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, pages 1396–1406, 2021.
9. Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *ECCV*, pages 265–282, 2024.
10. Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *ICCV*, pages 13869–13879, 2025.
11. Jinseok Bae, Inwoo Hwang, Young-Yoon Lee, Ziyu Guo, Joseph Liu, Yizhak Ben-Shabat, Young Min Kim, and Mubbasir Kapadia. Less is more: Improving motion diffusion models with sparse keyframes. In *ICCV*, pages 11069–11078, 2025.
12. Pengfei Zhang, Pinxin Liu, Pablo Garrido, Hyeongwoo Kim, and Bindita Chaudhuri. Kinmo: Kinematic-aware human motion understanding and generation. In *CVPR*, pages 11187–11197, 2025.
13. Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. *arXiv preprint arXiv:2505.01425*, 2025.

14. Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, and Daniel Cremers. Nonisotropic gaussian diffusion for realistic 3d human motion prediction. In *CVPR*, pages 1871–1882, 2025.
15. Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. Human motion instruction tuning. In *CVPR*, pages 17582–17591, 2025.
16. German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, Ferran Rigual, Nicky Sijia He, Cristina Palmero, Sergio Escalera, Yuting Ye, et al. From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In *CVPR*, pages 1850–1860, 2025.
17. Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *CVPR*, pages 22746–22755, 2025.
18. Dong Wei, Xiaoning Sun, Xizhan Gao, Shengxiang Hu, and Huaijiang Sun. Alien: Implicit neural representations for human motion prediction under arbitrary latency. In *CVPR*, pages 1861–1870, 2025.
19. Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. Lgtm: Local-to-global text-driven human motion diffusion model. In *SIGGRAPH*, 2024.
20. Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. In *NeurIPS*, pages 130739–130763, 2024.
21. Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *ECCV*, pages 37–54, 2024.
22. Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *ECCV*, pages 405–421, 2024.
23. Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Runyi Yu, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Local action-guided motion diffusion model for text-to-motion generation. In *ECCV*, pages 392–409, 2024.
24. Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *CVPR*, pages 1357–1366, 2024.
25. Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. In *ICLR*, 2025.
26. Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. In *CVPR*, pages 27849–27858, 2025.
27. Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. In *NeurIPS*, pages 28051–28077, 2024.
28. Yijun Qian, Jack Urbanek, Alexander Hauptmann, and Jungdam Won. Text motion translator: A bi-directional model for enhanced 3d human motion generation from open-vocabulary descriptions. In *ECCV*, pages 398–414, 2024.
29. Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. 2016.
30. Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H Bermano. Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion. In *CVPR*, pages 1965–1974, 2024.

31. Yuan Wang, Zhao Wang, Junhao Gong, Di Huang, Tong He, Wanli Ouyang, Jile Jiao, Xuetao Feng, Qi Dou, Shixiang Tang, et al. Holistic-motion2d: Scalable whole-body human motion generation in 2d space. *arXiv preprint arXiv:2406.11253*, 2024.
32. Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pages 14730–14740, 2023.
33. Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *ICLR*, 2025.
34. Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024.
35. Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, pages 9488–9497, 2023.
36. Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497. Springer, 2022.
37. Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, pages 390–408, 2024.
38. Yu Hua, Weiming Liu, Gui Xu, Yaqing Hou, Yew-Soon Ong, and Qiang Zhang. Deterministic-to-stochastic diverse latent feature mapping for human motion synthesis. In *CVPR*, pages 22724–22734, 2025.
39. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
40. Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*.
41. Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2022.
42. Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. Mixermdm: Learnable composition of human motion diffusion models. In *CVPR*, pages 12380–12390, 2025.
43. Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260*, 2024.
44. Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025.
45. Seunggeun Chi, Hyung-gun Chi, Hengbo Ma, Nakul Agarwal, Faizan Siddiqui, Karthik Ramani, and Kwonjoon Lee. M2d2m: Multi-motion generation from text with discrete diffusion models. In *ECCV*, pages 18–36, 2024.
46. Wenjie Zhuo, Fan Ma, and Hehe Fan. Infinidreamer: Arbitrarily long human motion generation via segment score distillation. In *ICCV*, pages 14688–14698, 2025.
47. Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. In *AAAI*, 2025.
48. Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *CVPR*, pages 17592–17602, 2025.

49. Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *ICCV*, pages 509–519, 2023.
50. Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, 2024.
51. Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: bidirectional autoregressive motion model. In *ECCV*, 2024.
52. Minjae Jeong, Yechan Hwang, Jaejin Lee, Sungyoon Jung, and Won Hwa Kim. Hgm^3: Hierarchical generative masked motion modeling with hard token mining. In *ICLR*, 2025.
53. Junyu Shi, Lijiang Liu, Yong Sun, Zhiyuan Zhang, Jinni Zhou, and Qiang Nie. Genm3: Generative pretrained multi-path motion model for text conditional human motion generation. *arXiv preprint arXiv:2503.14919*, 2025.
54. Yilin Wang, Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Xinxin Zuo, Juwei Lu, Hai Jiang, and Li Cheng. Motiondreamer: One-to-many motion synthesis with localized generative masked transformer. *arXiv preprint arXiv:2504.08959*, 2025.
55. Runqi Wang, Caoyuan Ma, Guopeng Li, Hanrui Xu, Yuke Li, and Zheng Wang. You think, you act: The new task of arbitrary text to motion generation. In *ICCV*, pages 12012–12022, 2025.
56. Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.
57. Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016.
58. Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015.
59. Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, pages 15085–15099, 2023.
60. Jiaman Li, C Karen Liu, and Jiajun Wu. Lifting motion to the 3d world via 2d diffusion. In *CVPR*, pages 17518–17528, 2025.
61. Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, pages 347–365, 2020.
62. Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *CVPR*, pages 3093–3102, 2023.
63. Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Luc Van Gool, and Dengxin Dai. Weakly supervised 3d object detection from lidar point cloud. In *ECCV*, pages 515–531, 2020.
64. Zhichao Zhai, Guikun Chen, Wenguan Wang, Dong Zheng, and Jun Xiao. Taga: Self-supervised learning for template-free animatable gaussian articulated model. In *CVPR*, pages 21159–21169, 2025.
65. Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.
66. 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 *ICML*, pages 8748–8763, 2021.

67. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.
68. Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
69. Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021.
70. Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020.
71. Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. Ms2sl: Multimodal spoken data-driven continuous sign language production. In *ACL*, pages 7241–7254, 2024.
72. Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Exploring vision transformers for 3d human motion-language models with motion patches. In *CVPR*, pages 937–946, 2024.
73. Zhe Li, Weihao Yuan, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, Laurence Tianruo Yang, et al. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. In *ICLR*, 2025.
74. Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021.
75. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.
76. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.
77. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
78. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
79. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
80. Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*, pages 16857–16867, 2020.
81. Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.
82. Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
83. Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, Lihong Liu, and Xingang Wang. Humandreamer: Generating controllable human-motion videos via decoupled generation. In *CVPR*, pages 12391–12401, 2025.
84. Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, pages 4210–4220, 2023.

85. Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *NeurIPS*, 2023.
86. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017.
87. Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.
88. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.
89. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
90. Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022.
91. Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM TOG*, pages 1–15, 2020.
92. Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arxiv:2310.12978*, 2023.