UniMotion: Unifying 3D Human Motion Synthesis and Understanding

Chuqiao Li [1] Julian Chibane [1,2] Yannan He [1] Naama Pearl [1]
Andreas Geiger [1] Gerard Pons-Moll[1,2]

[1]Tübingen AI Center, University of Tübingen, Germany, [2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

{chuqiao.li, yannan.he, naama.pearl, a.geiger, gerard.pons-moll}@uni-tuebingen.de,
jchibane@mpi-inf.mpg.de

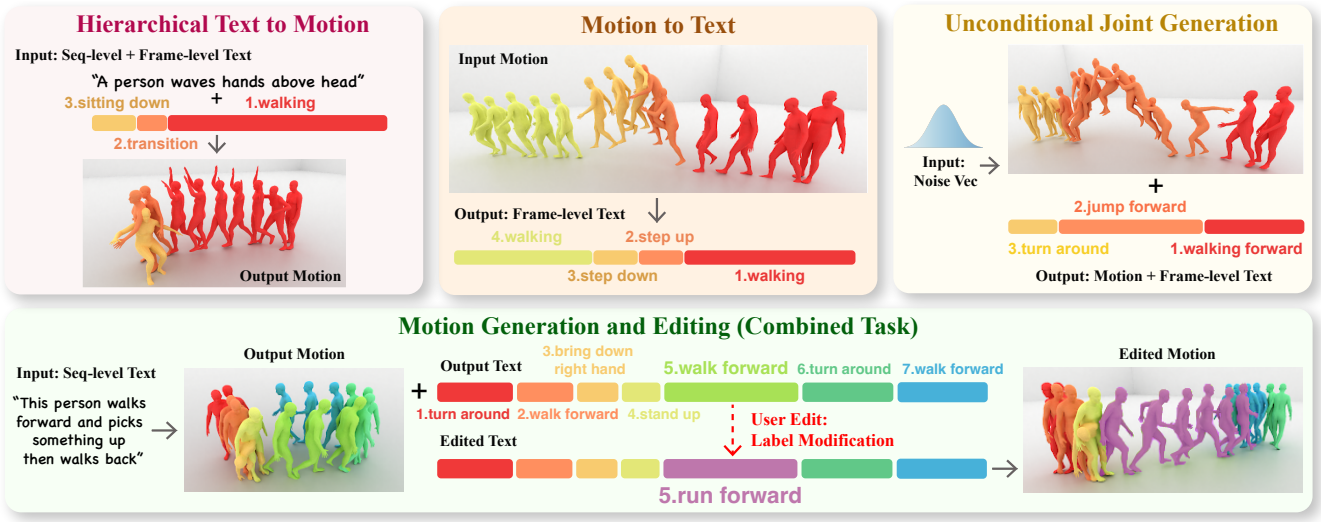

Figure 1. **Universality of UniMotion.** Our model can generate motion from compositional sequence- and frame-level text (**Hierachical Text to Motion**), generate detailed per frame motion descriptions (**Motion to Text**), generate motion and accurate frame-level text descriptions from noise (**Unconditional Joint Generation**), amongst other use cases outlined in our experiments section. Tasks can be combined for a controllable generation: users can generate motion from a coarse sentence, our model additionally generates detailed text descriptions, which can be edited and used for regeneration, generating the desired edited motion (**Motion Generation and Editing**).

Abstract

We introduce UniMotion, the first unified multi-task human motion model capable of both flexible motion control and frame-level motion understanding. While existing works control avatar motion with global text conditioning, or with fine-grained per frame scripts, none can do both at once. In addition, none of the existing works can output frame-level text paired with the generated poses. In contrast, UniMotion allows to control motion with global text, or local frame-level text, or both at once, providing more flexible control for users. Importantly, UniMotion is the first model which by design outputs local text paired with the generated poses, allowing users to know what motion happens and when, which is necessary for a wide range of applications. We show UniMotion opens up new applications: 1.) hierarchical control, allowing users to specify motion at different levels of detail, 2.) obtaining motion text descriptions for existing MoCap data or youtube videos 3.) allowing for editability, generating motion from text and editing the motion via text edits. Moreover, UniMotion attains state-of-the-art results for the frame-level text-to-motion task on the established HumanML3D dataset. The pre-trained model and code are available available on our project page at https://coral79.github.io/uni-motion/.

1. Introduction

Human motion synthesis is important for gaming, robotics and AR/VR applications. In real-world scenarios, avatars need to be controlled at multiple levels of abstraction. Effective controllability requires that an avatar be capable of executing detailed local sub-tasks according to a timeline while simultaneously understanding the overall global objective. In addition, the synthesis model should be aware of what action happens and when – an essential feature of biological intelligence to react to the external world.

However, current motion synthesis methods focus on either global per sequence-level control, or local per frame-level control, but don't allow for both. This results in single-level conditioning, thereby **lacking hierarchical control.** Importantly, these models also lack fine-grained motion awareness, specifically, the ability to output motion descrip-

tions for each pose in the generated output motion sequence. The frame-level text-to-motion methods [1, 3, 33] provide detailed manipulation of individual frames. However, it can be impractical to specify the exact duration of each action in some situations, and ensuring overall semantic plausibility throughout the entire sequence remains challenging for these models. Conversely, the sequence-level text-to-motion methods [17, 37, 51] focus on achieving natural overall motion but struggles with fine-grained control. Furthermore, current models lack semantic awareness of the synthesized motion – there is no understanding of what action occurs when. Thus, they are **lacking motion understanding**, which is crucial for reacting to the external world and allows for action-specific editing in animation applications. While some works have made progress in this direction [17, 51] by predicting sequence-level text descriptions from motion, they fail to provide fine-grained frame-level text. Overall, despite their potential synergies, motion understanding and synthesis have been treated in isolation in the literature.

In this paper, we introduce UniMotion, the first unified multi-task model capable of both flexible motion control and frame-level motion understanding. UniMotion takes as input, global sequence level or local frame-level text inputs or human motion sequences, or any subsets thereof, or no input in case of unconditional generation. The output of our model is either fine grained, per pose text descriptions, or human motion sequences. This flexibility, allows us to train our model from different data sources. Moreover, by design, we unify tasks that are usually treated in separation by prior works, such as *Frame-Level Text-to-Motion*, *Sequence-Level Text-to-Motion* and *Motion-to-Text*, into a single simple unified model, trained a single time. Importantly, UniMotion's flexibility also allows for novel tasks not previously considered by prior work like 1.) unconditional generation of human motion with corresponding frame-level text descriptions and 2.) generation of frame-level text from motion, providing granular, time-aware annotations (see Fig. 1 for an illustration our diverse tasks).

To accomplish this, our model utilizes a transformer architecture with temporal alignment between the motion and frame-level text. We further enhance this by diffusing the local text together with the poses, using different diffusion time variables for each, inspired by the approach in Unidiffuser [2]. Specifically, the local text is tokenized and frame-wise aligned with the 3D poses, while the global text is injected as a global token. This design allows UniMotion to dynamically switch between global, local, or combined conditioning signals at test time, providing flexibility in motion generation and understanding. During training, we sample from all possible distributions (global and/or local conditioning, or unconditional), alternating between providing noise and signal to the model for each modality. This method effectively teaches the model both unconditional and conditional distributions, equipping it with the ability to handle various inputs.

Real-world applicability. We demonstrate practical utility across various real-world scenarios:

2D Video Annotation: We annotate human motion extracted from YouTube videos with frame-level text, by feeding UniMotion with human pose estimation (HPE) results. This annotation can serve as close captions for the visually impaired. *4D Mocap Annotation:* We annotate human motion captures, e.g. obtained from IMUs, with frame-level text. This provides automated insights and descriptions into the captured motions, e.g. allowing for text search retrieval of motion sequences. *Hierachical Control:* We provide examples of generating motion sequences with two levels of abstraction, specifying a general motion for arms via global text, and a fine-grained motion sequence for the rest of the body via local-level text. *Motion editing:* We show that UniMotion can be used for content creation, where controllability of the motion is important. Given a global text description, a user can generate an initial motion including a local-level text description. The user can then edit the motion as desired, editing the text segments and regenerating the motion. In summary, our **key contributions** are:

- **Unified Synthesis and Understanding**: We introduce UniMotion, the first unified probabilistic motion model allowing for sampling from the joint and all possible conditionals. It unifies tasks that are usually treated in separation by prior works, while also allowing for novel tasks not previously considered.
- **Results and Applications**: We show applicability to 2D Video Annotation, 4D Mocap Annotation, Hierachical Control and Motion editing. Moreover UniMotion attains state-of-the-art results for the frame-level text-to-motion task on the established HumanML3D dataset. Code and models will be released upon acceptance.

2. Related Work

Conditional human motion synthesis.. Synthesizing human motion has been a long-standing challenge. Recent studies in motion generation have shown notable progress in synthesizing movements conditioned on diverse modalities such as text [26, 28, 33, 34, 36, 37], music [20, 21], scenes [25, 38], and interactive objects [13, 19, 35, 40, 42, 43, 48, 49]. Recent years have witnessed substantial advancements in text-driven motion generation [8, 11, 12, 26, 36, 44]. Notably, diffusion-based generative models have emerged as potent tools, exhibiting impressive performance on leading benchmarks for text-to-motion tasks. Pioneering efforts such as MotionDiffuse [45], MDM [37], and FLAME [18] represent early applications of diffusion models to text-driven motion generation. Building upon

this foundation, MLD [6] further harnesses latent diffusion models, while ReMoDiffuse [46] integrates retrieval techniques into the motion generation pipeline. Recent MotionLCM [8] accelerates the sampling speed by adopting consistency model in motion latent space. Noteworthy, OmniControl [41] specializes in fine-grained spatial control of body joints.

Text-to-motion generation models.. The current landscape of text-to-motion generation models can be categorized into two main streams of controllability: (a) global text-based control and (b) Fine-grained local text-based control. Among the former, MotionGPT [17] utilizes pretrained language models and motion-specific vector quantized models to conceptualize human motion as a language. Similarly, AvatarGPT [51] proposes a top-down approach to address end-to-end motion planning and synthesis.

Conversely, research focusing on short, specific instructions presents another avenue. PriorMDM [33] introduces a two-stage method that synthesizes short motion sequences and their padded transitions. However, due to the lack of effective supervised learning, motions generated by such methods often exhibit artifacts, such as abrupt speed changes. FineMoGen [47] proposes diffusion-based motion generation and editing for fine-grained per-body part motion control, albeit requiring detailed per-body part instructions as input. Closely aligned with our work are methods enabling temporal control of motion, where the length of each motion segment can be controlled at the frame level. FlowMDM [3] demonstrates impressive results in seamless transitions between local motion segments, while STMC [28] proposes a hybrid method for spatial and temporal motion composition of multi-stream motion using off-the-shelf pre-trained motion models [37]. Notably, these methods do not condition on global text, resulting in a lack of awareness of the global motion context and less natural motion transitions.

Our method combines the advantages of both categories. It is the first method enabling the generation of human motion conditioned both at the abstract level with global text and at the detailed level with local texts.

Human motion understanding.. Understanding the meaning of human motion has been a long-standing research topic, this has been approached by describing human motion with predefined action labels [7, 52], which have dominated this field for some time. However, these methods have obvious limitations, they are not appropriate to describe complex motion sequences. Recently, the text annotated motion datasets [5, 11, 29] have enabled the methods [12, 17, 44] that learn the mutual mapping between human motion sequences and natural language descriptions. While these works produce impressive, they fall short in generating accurate per-frame language descriptions. More recently, methods such as [9, 16] have achieve motion editing based on more fine-grained conditions, such as per body part condition. However, they still lack the capability for temporal editing. UniMotion is the first approach that not only generates per-frame language descriptions but also allows for motion generation over specified time spans, thus advancing the understanding of human motion.

3. Preliminary: Motion Diffusion Model

We provide a brief overview of the Human Motion Diffusion Model (MDM) [37], which is designed for sequence-level text-to-motion synthesis. This model serves as a building block for our UniMotion, which extends its capabilities by **(a)** incorporating frame-level text input and **(b)** enabling the joint generation of both motion and text. MDM aims to synthesize human motion sequences, denoted as $\mathbf{x}^{1:N}$, where N is the length of the sequence. The synthesis process is guided by a sequence-level text condition c, meaning the entire motion sequence is described by a single text prompt. In cases of unconditioned motion generation, the condition is represented as $c = \emptyset$.

Diffusion is modeled as a Markov noising process, where $t = 0$ represents the timestep corresponding to the clean data and $t = T$ corresponds to the fully corrupted data. The samples generated during this process are denoted as $\{\mathbf{x}_t^{1:N}\}_{t=0}^{T}$, with $\mathbf{x}_0^{1:N}$ being drawn from the data distribution. The transition between steps is defined by:

$$q(\mathbf{x}_t^{1:N}|\mathbf{x}_{t-1}^{1:N}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}^{1:N}, (1-\alpha_t)\mathbf{I}). \quad (1)$$

where $\alpha_t \in (0, 1)$ indicates the noise level, with $\alpha_i = 1 - \beta_i$ and β_i being the noise schedule. We drop the sequence length and use $\mathbf{x}_t$ to denote the full sequence at noising step t for simplicity. The reverse diffusion process gradually denoises the noisy sequence $\mathbf{x}_T$, with the conditioned motion generation modeling the distribution $p(\mathbf{x}_0|c)$. The denoised data is directly predicted using a model G, where $\hat{\mathbf{x}}_0 = G(\mathbf{x}_t, t, c)$ [32].

To adapt the diffusion model for human motion, we follow [11] to parameterize the human motion as a 263 dimensional vector. Due to its redundancy inherent in the motion representation, a simple training objective [15, 37] can be used, minimizing the expected distance between the original noisy motion $\mathbf{x}_0$ and the predicted motion $\hat{\mathbf{x}}_0$:

$$\mathcal{L}_{\text{simple}} = E_{\mathbf{x}_0 \sim q(\mathbf{x}|c), t \sim \mathcal{U}\{1,\dots,T\}} \|\mathbf{x}_0 - G(\mathbf{x}_t, t, c)\|_2^2. \quad (2)$$

Notably, this simple loss automatically includes the geometry losses terms described by [37], enforcing physical plausibility and preventing artifacts.

4. UniMotion: Unifying Motion Synthesis and Understanding

In this section, we introduce UniMotion, a unified model for joint motion synthesis and understanding, including hierar-

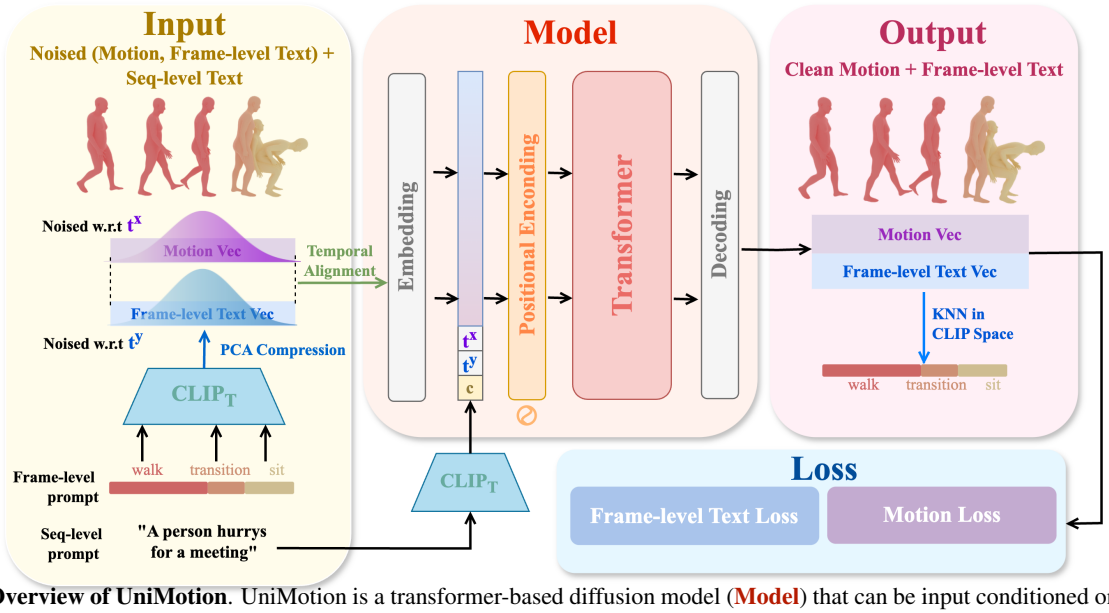

Figure 2. **Overview of UniMotion**. UniMotion is a transformer-based diffusion model (**Model**) that can be input conditioned on a) human motion, b) clip embedded frame-level text, or c) sequence-level text (**Input**) or any subsets thereof or none, and instead supplied with noise. At it's core it allows to diffuse motion and text individually, implemented via separate denoising timesteps t^x and t^y. After training with Frame-level text Losses and Motion losses (**Loss**), see Sec. 4.1. UniMotion can output clean, noise-free motion, and frame-level text descriptions explaining the generated motions. (**Output**)

chical control via text. UniMotion generates high-quality motion and text outputs, either from full noise or given conditional inputs such as frame-level text, sequence-level text, a motion sequence, or any subsets thereof, (see Fig. 2) spanning a variety of applications treated in isolation by related works.

To achieve this, our model advances prior single-modality motion diffusion models (see Subsec. 4.1) to encompass multi-modal distributions, specifically motion, and fine-grained text. We combine motion sequence and fine-grained frame-level texts, maintaining the temporal alignment of these two modalities to enable temporal semantic awareness (see Sec. 4.2). Unlike previous works, our multi-modality diffusion process supports joint training across datasets with varying annotations (sequence-level and frame-level) (see Subsec. 4.3).

4.1. Multi-Modal Motion and Text Diffusion

Previous motion synthesis models mainly focus on text-to-motion synthesis tasks [1, 3, 8, 17, 28, 33, 37, 41, 45, 47, 50, 51]. Some recent methods also generate sequence-level text descriptions [17, 51], but lack the temporal awareness and alignment we propose with UniMotion. Moreover, no model currently supports the joint generation of motion and text. This motivates our holistic model of motion synthesis and understanding, UniMotion, working in a multi-modal, joint probabilistic fashion we introduce next.

Similar in spirit to [2] that focuses on joint probabilistic modeling of 2D images and text, our method models the distribution of two temporal modalities under global conditioning. More concretely, a frame-level text sequence, $\mathbf{y}^{1:N}$ is denoted analogously to the motion sequence $\mathbf{x}^{1:N}$, where N denotes the sequence length and $\{\mathbf{y}_t^{1:N}\}_{t=0}^T$ are the noise samples created via Eq. 1. Similarly, we drop the notation of sequence length in the following for simplicity. With that, multi-modal diffusion can be achieved by extending G to $G_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}; t^x, t^y)$ with the additional process and including the separately scheduled diffusion timesteps t^x, t^y for motion and text respectively. By virtue of this formulation, the joint distribution $p(\mathbf{x}, \mathbf{y})$ can be sampled at inference time, starting the denoising process with $G_\theta(\mathbf{x}_T, \mathbf{y}_T; T, T)$, and the conditional $p(\mathbf{x}|\mathbf{y})$ by $G_\theta(\mathbf{x}_T, \mathbf{y}; T, 0)$ and analogously $p(\mathbf{y}|\mathbf{x})$. Specifically, we jointly train the model via

$$\min_\theta \mathbb{E}_{(\mathbf{x}_0,\mathbf{y}_0),t^x,t^y} \mathbb{E}_{\mathbf{x}_{t^x},\mathbf{y}_{t^y}} \|G_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}; t^x, t^y, c) - (\mathbf{x}_0, \mathbf{y}_0)\|_2^2 \tag{3}$$

where θ are weights parametrizing G and $\mathcal{U}$ is the discrete uniform distribution and expectation is taken over distributions: $(\mathbf{x}_0, \mathbf{y}_0) \sim p(x, y)$, $t^x \sim \mathcal{U}\{0, ..., T\}$, $t^y \sim \mathcal{U}\{0, ..., T\}$, $\mathbf{x}_{t^x} \sim q(\mathbf{x}_{t^x}|\mathbf{x}_0)$, $\mathbf{y}_{t^y} \sim q(\mathbf{y}_{t^y}|\mathbf{y}_0)$.

4.2. Temporally aligned Text and Motion Encoding

We find that appropriate architectural integration of two modalities (text and motion) into the joint formulation is a key performance factor.

Method	Training Set	Input	Per-crop semantic correctness			Per-crop Realism		Per-seq Realism	
			R-Prec@3 ↑	M2T ↑	M2M ↑	FID ↓	Diversity →	FID ↓	Diversity →
GT	-	-	$0.735^{\pm 0.008}$	$0.663^{\pm 0.000}$	$1.000^{\pm 0.000}$	$0.000^{\pm 0.000}$	$1.375^{\pm 0.005}$	$0.000^{\pm 0.000}$	$1.391^{\pm 0.003}$
TEACH	BABEL	f	$0.588^{\pm 0.007}$	$0.623^{\pm 0.001}$	$0.575^{\pm 0.000}$	$0.155^{\pm 0.001}$	$1.340^{\pm 0.003}$	$0.304^{\pm 0.001}$	$1.344^{\pm 0.003}$
DoubleTake	BABEL	f	$0.544^{\pm 0.013}$	$0.602^{\pm 0.002}$	$0.560^{\pm 0.001}$	$0.195^{\pm 0.002}$	$1.332^{\pm 0.005}$	$0.353^{\pm 0.002}$	$1.337^{\pm 0.004}$
STMC	HML	f	$0.528^{\pm 0.012}$	$0.599^{\pm 0.000}$	$0.616^{\pm 0.010}$	$0.156^{\pm 0.000}$	$1.358^{\pm 0.005}$	$0.233^{\pm 0.000}$	$1.362^{\pm 0.005}$
FlowMDM	BABEL	f	$0.618^{\pm 0.007}$	$0.631^{\pm 0.002}$	$0.652^{\pm 0.001}$	$0.101^{\pm 0.001}$	$1.352^{\pm 0.006}$	$0.211^{\pm 0.002}$	$1.375^{\pm 0.005}$
Ours	BABEL	f	$0.636^{\pm 0.017}$	$0.633^{\pm 0.004}$	$0.677^{\pm 0.002}$	$0.087^{\pm 0.002}$	$1.366^{\pm 0.009}$	$0.180^{\pm 0.004}$	$1.374^{\pm 0.002}$
Ours	HML∩ BABEL	f	$0.668^{\pm 0.009}$	$0.643^{\pm 0.002}$	$0.698^{\pm 0.002}$	$0.071^{\pm 0.001}$	$1.372^{\pm 0.005}$	$0.150^{\pm 0.001}$	$1.378^{\pm 0.003}$
Ours	HML∩ BABEL	f + s	$\mathbf{0.679}^{\pm 0.006}$	$\mathbf{0.644}^{\pm 0.001}$	$\mathbf{0.706}^{\pm 0.002}$	$\mathbf{0.066}^{\pm 0.002}$	$\mathbf{1.373}^{\pm 0.009}$	$\mathbf{0.133}^{\pm 0.004}$	$\mathbf{1.381}^{\pm 0.006}$

Table 1. **Frame-Level to Text evaluation.** *Per-crop* refers to text segment level evaluation. *Training Set* specifies the dataset used for training. *Input* specifies the type of text input. *f*: frame-level text, *s*: sequence-level text. *f+s* demonstrates that combining multi-level conditioning signals can enhance model performance in terms of semantic correspondence. The evaluation is repeated 10 times, and ± indicates the 95% confidence intervals.

A simple integration is to treat motion and text as separate modalities and as input to the Transformer. An even more structure-agnostic approach is proposed by UniDiffuser [2], where each token of their text encoding is fed separately as input. We find that both these variants lead to performance issues.

In contrast, in our setting of motion and text sequences, we find temporal alignment to be the key. A simple, yet effective implementation is the concatenation of motion and text into joint encodings along the temporal dimension. Instead of learning to correlate word positions with motion positions, alignment is directly given through the input encoding.

However, this alone does not guarantee performance. We encode text into the space of CLIP [31] with a pertained model. Using the full encodings of pose and text as token creates issues. We hypothesize this is due to an excessive capacity spent on the high-dimensional text tokens. We solve this by projecting CLIP embeddings down to 50 dimensions via PCA [39] and find this improves performance drastically. To get back to text labels from embeddings after diffusion, we match the predicted clip embedding to our database of text labels to obtain the output text using the closest match.

4.3. Data Merging

The popular AMASS dataset [23] of natural human motion, represented by the SMPL body model [22] has recently been annotated in two efforts, namely BABEL [29] and HumanML3D [11]. While HumanML3D annotations consist of sequence-level text annotation, that is, a single text annotation for a motion clip, the BABEL annotations consist of frame-level annotations, assigning semantic label to the pose for each frame of the motion sequence. Instead of restricting to use one at a time, as in prior works, UniMotion is directly trained on both jointly, using sequence level HumanML3D annotations as condition c and frame level sequences as $\mathbf{y}^{1:N}$.

A challenge however lies in that both datasets annotate different subsets of AMASS. A trivial solution is to consider overlapping annotations of motions. We denote our model trained with this scenario UniMotion *overlap*, and investigate the performance in experiments (see Sec. 5).

5. Experiments

In this section we investigate the benefit of hierarchical text at inference and training time (i.e. usage of frame-level and sequence-level text). We show the versatility of UniMotion's unification of synthesis and understanding. Specifically, allowing for frame-level text to motion (Subsec. 5.1) and we for the first time show motion-to-frame-level text (Subsec. 5.2), including a real-world application scenario. Finally, UniMotion is the first model to show joint generation of motion along with frame-level understanding (Subsec. 5.2). In the ablation study, we show that the proposed multi-modality strongly improves generation quality compared to our backbone MDM [37]. (Subsec. 5.3)

Implementation Details.. We utilize a temporally aware transformer, similar to MDM [37]. Text inputs are encoded using pretrained CLIP, followed by PCA reduction. Our model is trained on a single A100, with training spanning approximately 40 hours. Please refer to 4.3 for details on training data.

Baselines. We compare our model to the publicly released works that are capable of frame-level text-to-motion generation: auto-regressive model **TEACH** [1], **DoubleTake** [33] based on diffusion sampling, **FlowMDM** [3], a diffusion model based on Blended Positional Encoding and **STMC** [28], a post-hoc test time method stichting individual predictions of MDM [37]. Note that neither Teach, FlowMDM nor STMC supports hierarchical training. Since STMC admits overlapping control signals we compare to it in terms of hierarchical control. Since no prior works allow for training on sequence and frame-level text input, models are either trained on BABEL(frame-level) or HumanML3D(sequence-level) data, as indicated in our result tables. Please refer to our supp. document for more

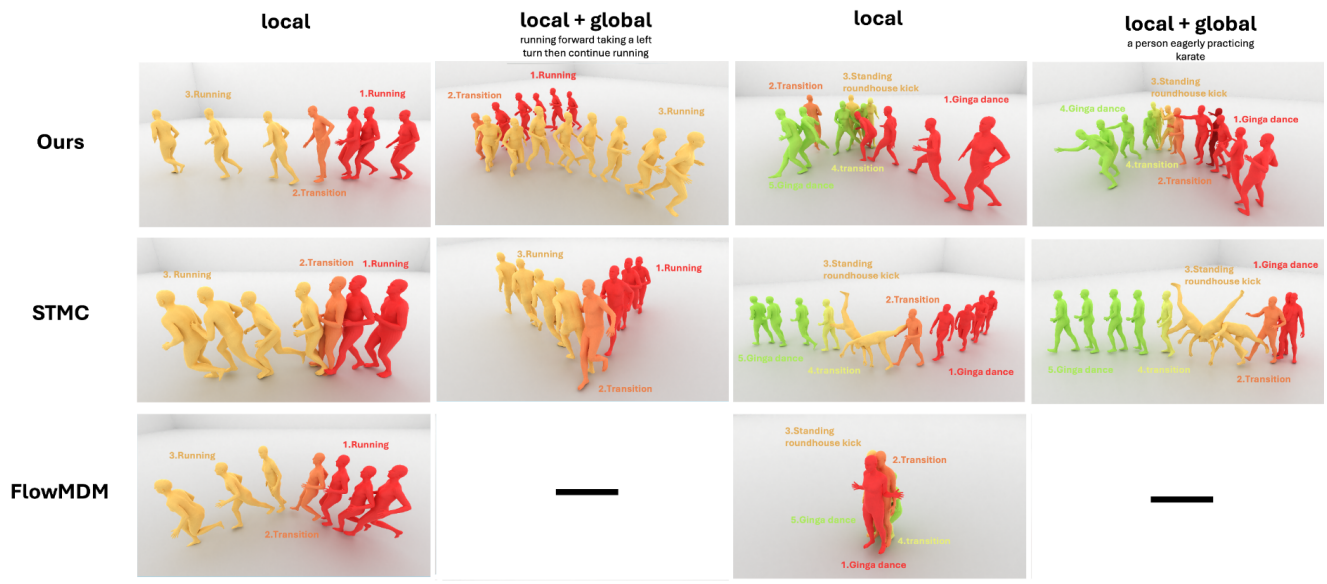

Figure 3. **Text2Motion qualitative results. Columns 1,3**: Local text is the input to our method and baselines STMC [28] (adapted) FlowMDM [3]. **Columns 2, 4**: Both local and global text are the input our method and STMC. Our model performs well regardless of the complexity of the local text, in contrast to STMC which fails to generate Ginga dance in columns 3 and 4 and performs walking instead. FlowMDM cannot be conditioned on both global+local text.

details.

Evaluation Metrics. First, we introduce our **semantic metrics**, measuring how well the generated motions correspond to their text descriptions. **R-Precision** [11] assesses the accuracy of ranking the correct ground-truth text corresponding to a predicted motion at the top positions (Top-1, Top-2, and Top-3) within a set that includes 32 randomly sampled incorrect text matches. With **M2T** [28], we measure how well the per-crop motion matches their textual description, we calculate their cosine similarity in the joint-embedding space of TMR++ [4]. Similarly, the **M2M** [28] score is the cosine similarity between the generated and the ground-truth motion embeddings.

With our **realism metrics**, we measure how well the generated motion distributions fit the ground truth one. We utilize the **Frechet Inception Distance (FID)** [14] to measure the distribution distances and **Diversity** [11] computes the distributions variance, both in the TMR++ as the embedding space.

5.1. Frame-Level Text2Motion Results

We evaluate the Text2Motion task (Tab. 1), where we investigate the effects of frame-level and sequence-level training data. Qualitative analysis is presented in Fig. 3. When we train our model as FlowMDM (best performing prior work) on frame-level labels of all Babel annotations (Tab. 1, Ours *BABEL*) we observe our UniMotion to be consistently better but still roughly on par as expected since both models are using a backbone similar to MDM [37]. The slight improvement can be attributed to the temporal input alignment (see Sec.4.2) and the multi-timestap diffusion training (see Sec. 4.1). Next, we significantly reduce the training dataset size to the subset sequences annotated with both HML (frame-level text) and BABEL (sequence-level text) (cf. Tab. 1 Ours HML-BABEL f). Although one could expect a performance decrease, we find the opposite, a strong consistent performance increase in all metrics - suggesting the strong positive impact of multi-model training. Notably, this is the case although only frame level inputs are given for the evaluation and sequence-level inputs only enrich the models training data. Finally, we investigate the effect of adding sequence-level text into the model for evaluation (cf. Tab. 1 Ours HML-BABEL f + s), again showing a consistent improvement. In conclusion, the evaluation shows cross-modal generalization, consistently improving the results.

5.2. Applications

Please see these and further results in motion in the supplementary video.

Motion2Text. Here we show UniMotions capabilities of predicting frame-level text given human motion. This is a novel task, prior work is not able to do. We, therefore, restrict ourselves to qualitative evaluations. See Fig. 4, where we use UniMotion to annotate MoCap data and Youtube videos with motion descriptions.

Hierarchical Text2Motion:. We show that UniMotion, although not directly trained for this task, shows generalization capabilities to compositional text conditioning, where

Method	Training Set	Input	FID ↓	Diversity →	R-Prec@1 ↑	R-Prec@2 ↑	R-Prec@3 ↑	M2T ↑
GT	-	-	$0.000^{\pm 0.000}$	$1.391^{\pm 0.003}$	$0.699^{\pm 0.014}$	$0.834^{\pm 0.011}$	$0.878^{\pm 0.005}$	$0.748^{\pm 0.000}$
MDM	HML	s	$0.449^{\pm 0.025}$	$1.315^{\pm 0.014}$	$0.376^{\pm 0.008}$	$0.536^{\pm 0.010}$	$0.639^{\pm 0.010}$	$0.631^{\pm 0.003}$
Ours	HML-BABEL	f	$0.152^{\pm 0.002}$	$1.377^{\pm 0.006}$	$0.344^{\pm 0.010}$	$0.508^{\pm 0.019}$	$0.587^{\pm 0.007}$	$0.648^{\pm 0.003}$
Ours	HML-BABEL	s	$0.195^{\pm 0.003}$	$1.381^{\pm 0.011}$	$0.375^{\pm 0.021}$	$0.539^{\pm 0.018}$	$0.655^{\pm 0.016}$	$0.653^{\pm 0.004}$
Ours	HML-BABEL	f + s	$\mathbf{0.133}^{\pm 0.003}$	$\mathbf{1.382}^{\pm 0.002}$	$\mathbf{0.424}^{\pm 0.005}$	$\mathbf{0.593}^{\pm 0.011}$	$\mathbf{0.677}^{\pm 0.011}$	$\mathbf{0.678}^{\pm 0.002}$

Table 2. **Ablation Study on Sequence-level Text2Motion generation.** In this table, we compare with our backbone model MDM[37] to study whether introducing multi-modality helps the motion generation performance. Symbols ↓, and → indicate that lower, or values closer to the ground truth (GT) are better, respectively. The evaluation is repeated 10 times, and ± indicates the 95% confidence interval.

Figure 4. **Motion2Text understanding of MoCap and YouTube data. (a)** Given an input MoCap sequence, we use UniMotion to predict frame-level local text. **(b)** We annotate human motion from YouTube videos with frame-level text. We lift 2D videos to 3D human motion via frame-by-frame pose estimators [10]. We visualize the SMPL human pose (**Pink**) overlayed on the YouTube videos frames. Then we run UniMotion to predict frame-level annotations (colored text descriptions below the frames). Annotations could serve as valuable audio close captions for the visually impaired.

global-text and local-text are giving different but complementary conditioning (see. Fig. 1).

Joint text and motion generation. UniMotion can jointly generate human motion and corresponding frame-level text, allowing users to not only generate motion but also to directly understand the generated sequence on a frame level. Prior work can not perform this task, see Fig. 5 for conditional joint generation and in Fig. 6 for unconditional generation.

Motion Editing for Content Creation. We show the application of UniMotion to content creation, where a user specifies a desired motion sequence via rough global text and obtains the motion sequence with a frame-level script. The user succeeds by editing the frame-level script and regenerates the motion to obtain the desired edits, see Fig. 1.

5.3. Ablation: Importance of Multi-Modality

In this section, we investigate the importance of our unification of multiple modalities.

Flexibility.. As seen in previous experiments, this allows to generate high-quality motion and text outputs, either from full noise or given conditional inputs such as frame-level text, sequence-level text, a motion sequence, or any subsets thereof, (see Fig. 2) spanning a variety of applications treated in isolation by related works.

Improved quality.. Additionally, we ablate that the included multi-modality also allows for improved generation quality. For this, we compare our model trained on multi-modal against our backbone architecture MDM [37], which does not include frame-level text in output or input, nor is equipped with the flexible multi-modal diffusion.

Our model, used with the same sequence-level text input data (Table 2, input: s), as MDM, drastically improves MDM in terms of FID and diversity, but also improves or

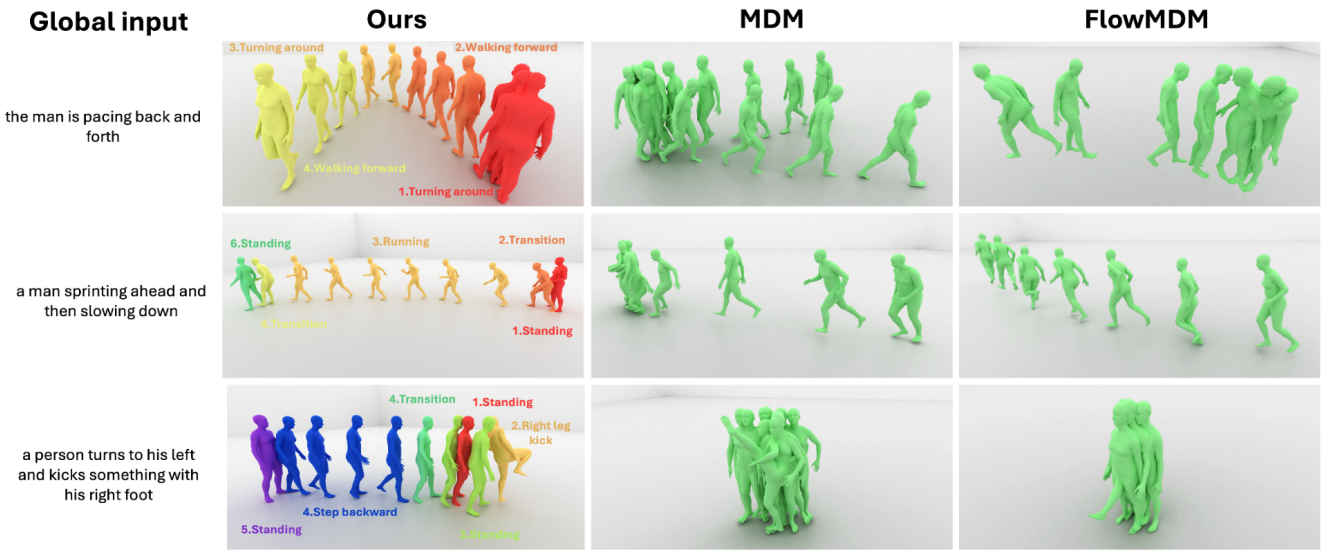

Figure 5. **Joint text and motion generation results.** Input to the models is only the global text shown on the left. We compare the generated motion of ours, MDM [37] and FlowMDM [3]. Our method jointly predicts the frame-level labels, so we can annotate subsequences, while MDM and FlowMDM can only generate the motion.

Figure 6. **Unconditional joint text and motion generation**. Our model, by design, generates poses aligned with local text.

is on par in other metrics. Since the backbone transformer is the same, this shows the strength of the proposed multi-modal training. Notably, this effect is visible even though our training dataset is only a 30% subset of the MDM training dataset.

Combining sequence-level and frame-level text (Table 2, input: f+s) shows a further significant improvement, improving MDM in all metrics. This improvement does not stem from the addition of frame-level text input alone since, in isolation, frame-level labels do not achieve this quality (see Table 2, input: f). We find the interaction between frame-level and sequence-level inputs is the reason for the improvements. In conclusion, the proposed multi-modality is the key factor allowing for improved generation quality.

6. Conclusions

We introduced UniMotion, the first unified multi-task human motion model capable of both flexible motion control and frame-level motion understanding. Using a flexible multi-model diffusion scheme, UniMotion solves several tasks in a unified fashion. Specifically, it unifies tasks that are usually treated in separation by prior works, such as *Frame-Level Text-to-Motion*, *Sequence-Level Text-to-Motion* and *Motion-to-Text*, into a single simple unified model, trained a single time. Importantly, UniMotion's flexibility also allows for novel tasks not previously considered by prior work like 1.) unconditional generation of human motion with corresponding frame-level text descriptions and 2.) generation of frame-level text from motion, providing granular, time-aware annotations. We show UniMotion opens up new applications: 1.) hierarchical control, allowing users to specify motion at different levels of detail, 2.) obtaining motion text descriptions for existing MoCap data or YouTube videos and 3.) allowing for editability, generating motion from text, and editing the motion via text edits. Moreover, UniMotion attained state-of-the-art results for the frame-level text-to-motion task on the established HumanML3D dataset showing the proposed multi-modality is the key factor allowing for improved generation quality.

Acknowledgments: Special thanks to Xiaohan Zhang for helping with the related work and other RVH and AVG members for the help and discussion. Thanks to Mathis Petrovich, Léore Bensabath, and Prof. Gül Varol for the discussion and helpful information on TMR++[4]. Prof. Gerard Pons-Moll and Prof. Andreas Geiger are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 - Project number 390727645. Gerard Pons-moll is endowed by the Carl Zeiss Foundation. Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533). Julian Chibane is a fellow of the Meta Research PhD Fellowship Program - area: AR/VR Human Understanding.

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 2, 4, 5, 1

[2] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2, 4, 5

[3] German Barquero, Sergio Escalera, and Cristina Palmero. Flowmdm: Seamless human motion composition with blended positional encodings. *arXiv preprint arXiv:2402.15509*, 2024. 2, 3, 4, 5, 6, 8

[4] Léore Bensabath, Mathis Petrovich, and Gul Varol. A cross-dataset study for text-based 3d human motion retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1932–1940, 2024. 6, 9, 3

[5] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 3

[6] 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. 3

[7] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[8] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024. 2, 3, 4

[9] Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 3

[10] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 7

[11] 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 (CVPR)*, 2022. 2, 3, 5, 6, 1

[12] 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. 2, 3

[13] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, 2021. 2

[14] 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 *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[16] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing, 2024. 3

[17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4, 5, 6

[18] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis and editing. *arXiv preprint arXiv:2209.00349*, 2022. 2

[19] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6), 2023. 2

[20] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 2

[21] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 2

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34, 2015. 5

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 5

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1

[25] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024. 2

[26] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[27] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[28] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Stmc: Multi-track timeline control for text-driven 3d human motion generation. *arXiv preprint arXiv:2401.08559*, 2024. 2, 3, 4, 5, 6

[29] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5

[30] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 5

[32] 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. 3

[33] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Priormdm: Human motion diffusion as a generative prior. In *ICLR*, 2023. 2, 3, 4, 5, 1

[34] Yi Shi, Jingbo Wang, Xuekun Jiang, and Bo Dai. Controllable motion diffusion model. *CoRR*, abs/2306.00416. 2

[35] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6), 2019. 2

[36] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[37] 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. 2, 3, 4, 5, 6, 7, 8

[38] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE. 2

[39] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists. 5

[40] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *Arxiv*, 2023. 2

[41] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4

[42] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2

[43] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. *arXiv:2404.10685*, 2024. 2

[44] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[45] 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. 2, 4

[46] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023. 3

[47] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS*, 36, 2024. 3, 4

[48] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[49] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. 2024. 2

[50] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv preprint arXiv:2403.07487*, 2024. 4

[51] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding, planning, generation and beyond. *CoRR*, abs/2311.16468, 2023. 2, 3, 4

[52] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

UniMotion: Unifying 3D Human Motion Synthesis and Understanding

Supplementary Material

In the following, we start with the supplementary video in Sec. A and discuss the details of training data in Sec. B. Then, we present the details of our evaluation setup in Sec. C, followed by implementation details in Sec. D, additional results in Sec. E and Sec. F. Finally, we demonstrate our model's advantage over LLMs and other motion-to-text models in Sec. G.

A. Video with Qualitative Results

We provide videos to further explain our method and to present the results with animated motions, showing a clearer comparison across various tasks and against other baselines. Supplementary results can be found in the accompanying ZIP file.

B. Training Data

UniMotion is trained on an overlapping subset of BABEL [30] and HumanML3D [11], utilizing both sequence-level and frame-level text as input. Fig. 7 illustrates the data alignment and merging process. However, since these two datasets are independently labeled and cover different subsets of AMASS [24], they do not fully overlap. The overlapping portion comprises only 8,829 motion sequences (excluding left-right flipping), which represents approximately 30.25% of the HumanML3D dataset (23,384 sequences). This overlapped dataset includes motion sequences, sequence-level text descriptions, and frame-level text descriptions.

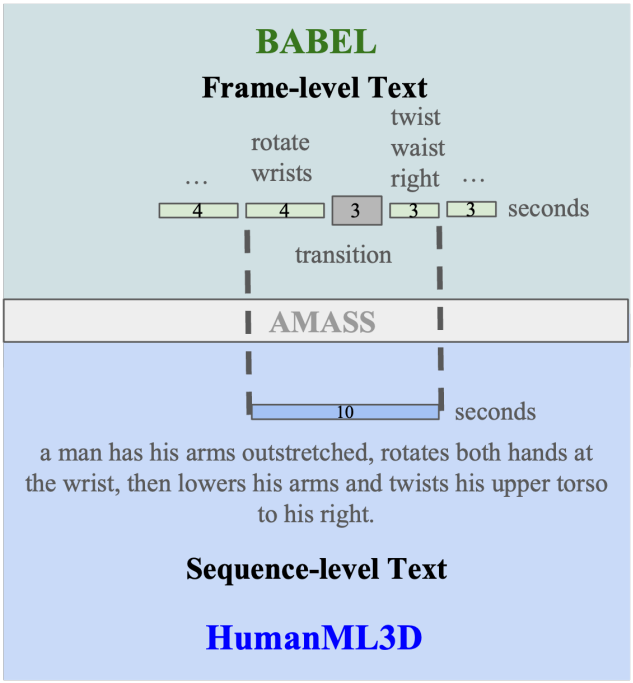

Figure 7. **We merged HumanML3D and BABEL based on their time correspondence with AMASS.** Each sequence (approximately 1-10 seconds) in HumanML3D includes 3-4 sequence-level annotations in sentence format, as illustrated in the blue area. In contrast, BABEL provides separate annotations for atomic actions with varying lengths, where the text labels are primarily short phrases aligned at the frame level, as shown in the green area.

C. Evaluation Setup

In this section, we outline the details of the evaluation setup and how we run baselines under this setup.

For frame-level text-to-motion generation, we use BABEL frame-level text (in short-phrase format) as conditional input, which is also used as our test-time text input. To ensure a fair comparison with other baselines and to maintain consistency with the training data distribution, we use their pre-trained models on BABEL if available. However, our model is trained on a subset of the HumanML3D training split, which overlaps with the BABEL test split. Consequently, we generate a joint test set, excluding training sequences from both. Finally, the test set contains 358 sequences and 998 sub-sequences of motion segments. Our test, train, and validation split will be made available alongside our code and models upon publication.

TEACH. For TEACH [1] we use the pre-trained model supplied by the authors on their website, which was trained on BABEL. Since TEACH can not be applied to text segments with very few frames, we set the minimum size of each evaluation sequence to 8 frames.

PriorMDM. For PriorMDM [33], we compare DoubleTake with our method. To fairly compare DoubleTake with our method, we use the "Babel_TransEmb_GeoLoss" pre-trained model, as our local text input is based on the BABEL dataset. When feeding motion crops into DoubleTake, we specify the length of each motion crop. In DoubleTake's default setup, the handshake size is set to 20 and the blending window size to 10, resulting in a minimum motion crop length of 70. If a motion crop is shorter than 70, the method automatically pads it to this length. However, many motion crops in our test set are shorter than 70, which would cause significant discrepancies between the input and output motion lengths. To maintain similar input and output sizes, we modify the handshake size to 2 and the blending window size to 1. The results under this setup are shown in Table 1.

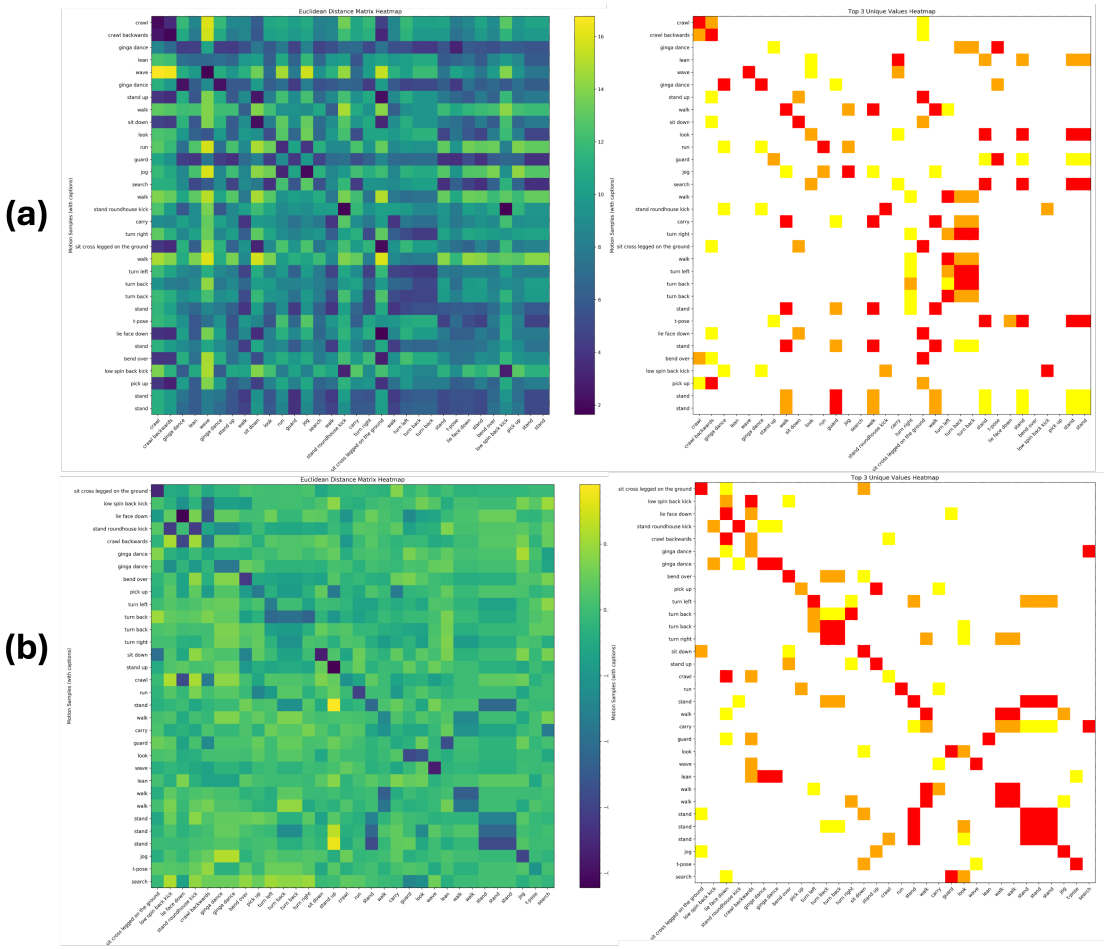

Figure 8. **The comparison between ground-truth motion-text matching** in the joint embedding spaces of Guo et al.'s model **(a)** and TMR++ **(b)**. **Left:** The heatmap shows the paired motion-text distances, where darker shades indicate smaller distances. The vertical axis represents motion samples, while the horizontal axis represents text samples. **Right:** The top-3 R-precision scores are displayed for each row, indicating the closest 3 texts to each motion. Red denotes the top 1 match, orange the top 2, and yellow the top 3. If the texts are identical, they are only counted as one.

STMC. For an entire motion sequence, STMC [28] allows specifying the body part for each individual subsequence of motion. To align with our setup, we set the corresponding body part to include all body parts for each motion crop when feeding the motion into STMC.

FlowMDM. To ensure a fair comparison with our method, we use the human motion compositions with the pre-trained BABEL model for the FlowMDM [3] method. Since FlowMDM is designed to generate motion compositions seamlessly, there is no need to specify any transition length between atomic motions. Therefore, we directly input the frame-level texts and corresponding lengths, consistent with the input format used for our model.

Evaluation metrics.. For the evaluation metrics—**Semantic Correspondence** (R-precision, M2T score, M2M score) and **Realism** (FID, Diversity)—we use TMR++ instead of the commonly used motion and text embedding model from Guo et al. [11]. This choice is driven by the need to evaluate models trained across different datasets and to assess performance at multiple levels of generated motion (per-crop vs. per-sequence).

For per-crop semantic correctness, we focus on evaluating the alignment of atomic motion crops with their corresponding input text, formatted as BABEL. Additionally, we assess the overall realism of sequence-level motion across crops, which aligns with HumanML3D's sequence-level evaluation. The evaluation model aims to establish a joint latent space for motion and text, performing matching between them based on distance within this shared space.

The commonly used model from Guo et al. [11] is trained solely on HumanML3D. To evaluate BABEL pre-trained models, Shafir et al. [33] retrained this model on BABEL data, and FlowMDM relies on these models

for separate evaluations on each dataset. STMC utilizes TMR [27], a retrieval model that demonstrates a better joint latent space compared to the classic evaluation model used by MDM, especially in terms of text-motion distance for ground-truth motion-text pairs. However, TMR is also trained only on HumanML3D, which limits its ability to accurately evaluate both crop-level motions and BABEL text, as well as sequence-level realism.

To address these limitations, we employ the latest model, TMR++ [4], which is trained across datasets and delivers highly accurate matching results between ground-truth motion and text, whether in BABEL format (subsequence level, short text phrases) or HumanML3D format (sequence-level, text descriptions in sentences).

For a quantitative comparison, please refer to Table 3, which evaluates ground-truth motion and text. For qualitative analysis, see Fig. 8, which presents a heatmap of the matching distance across a random sample of 32 batches.

Method	Training Set	Per-crop semantic correctness		
		R-Prec@1 ↑	R-Prec@2 ↑	R-Prec@3 ↑
Guo et al[11]	HumanML3D	$0.281^{\pm 0.005}$	$0.438^{\pm 0.004}$	$0.539^{\pm 0.006}$
TMR++[4]	HumanML3D+BABEL	$\mathbf{0.520}^{\pm 0.013}$	$\mathbf{0.659}^{\pm 0.008}$	$\mathbf{0.735}^{\pm 0.008}$

Table 3. **Ground-truth matching score comparison across evaluation modals.** In this table, we compare the matching scores across different evaluation models for ground-truth motion and text, averaging over batches of 32 random samples. The results demonstrate that TMR++ is a more reliable model within our evaluation setup.

D. Implementation Details

We provide more details about the implementation of our model. We extend the MDM [37] framework to separate time steps for motion and frame-level text, and adjust the input to accept the temporal alignment of both the motion vector and text embedding vector. The model is retrained from scratch using the merged overlapping dataset, with hyperparameters consistent with those suggested by Tevet et al. [37].

For frame-level text, we use the same CLIP model as used in MDM to generate embeddings. We then applied PCA to condense the dimensionality from 256 to 51, preserving approximately 70% of the original variance. Our model predicts both the clean motion and the condensed CLIP embeddings for the frame-level texts. To output the texts, we use K-nearest neighbors (KNN) to match the output CLIP embeddings in a pre-computed database. This approach effectively matches nearby CLIP embeddings to the corresponding closest text even with a small variance.

For the training and sampling algorithm, please refer to Algorithm 1, 2, 3.

Algorithm 1 Training

repeat
  $\mathbf{x}_0, \mathbf{y}_0, c \sim q(\mathbf{x}_0, \mathbf{y}_0, c)$
  $c = \varnothing$ with probability 10%
  $t^x, t^y \sim \text{Uniform}(\{1, 2, \dots, T\})$
  $\epsilon^x, \epsilon^y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Let $\mathbf{x}_{t^x} = \sqrt{\overline{\alpha}_{t^x}}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_{t^x}}\epsilon^x$
  Let $\mathbf{y}_{t^y} = \sqrt{\overline{\alpha}_{t^y}}\mathbf{y}_0 + \sqrt{1-\overline{\alpha}_{t^y}}\epsilon^y$
  Take gradient step on $\nabla_\theta \|\epsilon_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}, t^x, t^y, c) - [\mathbf{x}_0, \mathbf{y}_0]\|_2^2$
until converged

Algorithm 2 Sampling $\mathbf{x}_0$ conditioned on $\mathbf{y}_0$ (similar for sampling $\mathbf{y}_0$ conditioned on $\mathbf{x}_0$, with or without conditioning on c.

$\mathbf{x}_0^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$c = \varnothing$ or user specify
for $t = T, \dots, 1$ **do**
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_0^{t-1} = \epsilon_\theta^x(\sqrt{\overline{\alpha}_{t^x}}\mathbf{x}_0^t + \sqrt{1-\overline{\alpha}_{t^x}}\epsilon, \mathbf{y}_0, t, 0, c)$
end for
return $\mathbf{x}_0$

Algorithm 3 Joint sampling of $\mathbf{x}_0, \mathbf{y}_0$ (with or without condition on c)

$\mathbf{x}_0^T, \mathbf{y}_0^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$c = \varnothing$ or user specify
for $t = T, \dots, 1$ **do**
  $\epsilon^x, \epsilon^y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_0^{t-1}, \mathbf{y}_0^{t-1} = \epsilon_\theta(\sqrt{\overline{\alpha}_{t^x}}\mathbf{x}_0^t + \sqrt{1-\overline{\alpha}_{t^x}}\epsilon^x, \sqrt{\overline{\alpha}_{t^y}}\mathbf{y}_0^t + \sqrt{1-\overline{\alpha}_{t^y}}\epsilon^y, t, t, c)$
end for
return $\mathbf{x}_0, \mathbf{y}_0$

E. More Experiment Results

We provide only a subset of the metrics for semantic correspondence and motion realism in the main paper due to space constraints. Here, we provide the complete evaluation.

Semantic correspondence.. Tab. 4 lists all three R-precision scores, demonstrating that our method outperforms all baseline methods. These results are consistent with our conclusions in the experiment section of the main paper.

Realism. Tab. 5 includes FID and Diversity scores calculated using the evaluation model from Guo et al. [11] for reference. Note that at the crop level, this model provides less stable evaluations because it was trained only on

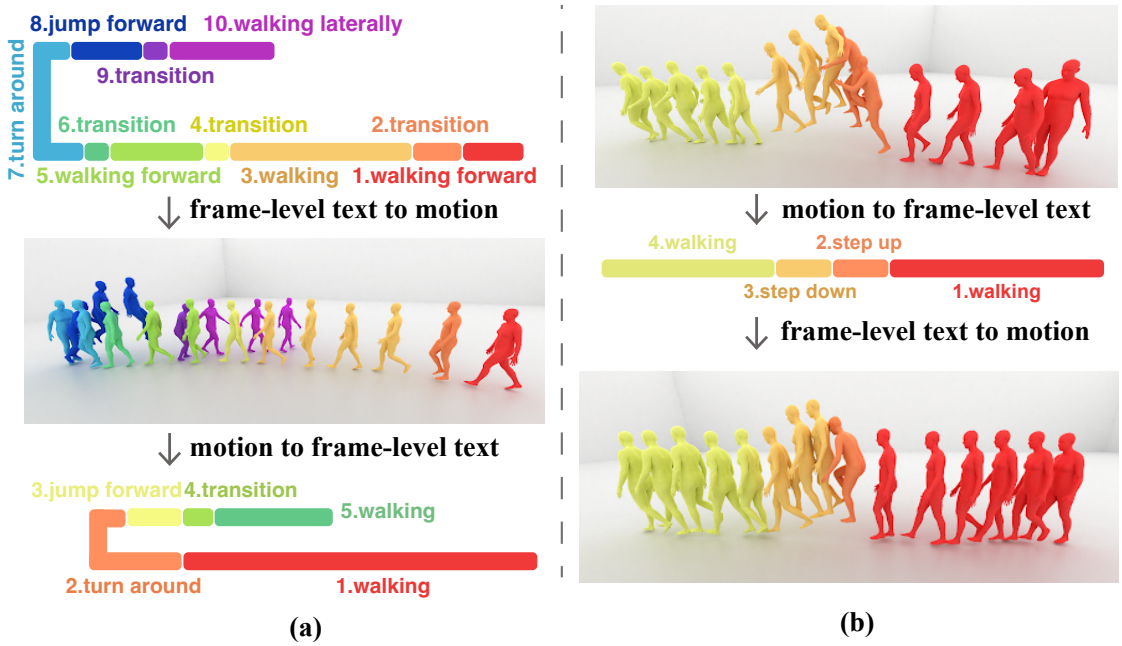

Figure 9. **Text variation (a) and motion variation (b)** are direct applications that leverage the two conditional distributions modeled by UniMotion. Motion variation **(b)** is achieved by generating frame-level text descriptions from a motion sequence, and then using these descriptions to create a new, semantically similar motion with different content. Text variation **(a)** is produced by reversing this process to create diverse text annotations.

Method	Training Set	Input	Per-crop semantic correctness				
			R-Prec@1 ↑	R-Prec@2 ↑	R-Prec@3 ↑	M2T ↑	M2M ↑
GT	-	-	$0.520^{\pm 0.013}$	$0.659^{\pm 0.008}$	$0.735^{\pm 0.008}$	$0.663^{\pm 0.000}$	$1.000^{\pm 0.000}$
TEACH	BABEL	f	$0.375^{\pm 0.008}$	$0.516^{\pm 0.007}$	$0.588^{\pm 0.007}$	$0.623^{\pm 0.001}$	$0.575^{\pm 0.000}$
DoubleTake	BABEL	f	$0.332^{\pm 0.013}$	$0.467^{\pm 0.013}$	$0.544^{\pm 0.013}$	$0.602^{\pm 0.002}$	$0.560^{\pm 0.001}$
STMC	HML	f	$0.321^{\pm 0.009}$	$0.452^{\pm 0.012}$	$0.528^{\pm 0.012}$	$0.599^{\pm 0.000}$	$0.616^{\pm 0.010}$
FlowMDM	BABEL	f	$0.389^{\pm 0.009}$	$0.532^{\pm 0.014}$	$0.618^{\pm 0.007}$	$0.631^{\pm 0.002}$	$0.652^{\pm 0.001}$
Ours	BABEL	f	$0.394^{\pm 0.010}$	$0.552^{\pm 0.018}$	$0.636^{\pm 0.017}$	$0.633^{\pm 0.004}$	$0.677^{\pm 0.002}$
Ours	HML-BABEL	f	$\underline{0.427}^{\pm 0.011}$	$\underline{0.587}^{\pm 0.012}$	$\underline{0.668}^{\pm 0.009}$	$\underline{0.643}^{\pm 0.002}$	$\underline{0.698}^{\pm 0.002}$
Ours	HML-BABEL	f + s	$\mathbf{0.450}^{\pm 0.018}$	$\mathbf{0.593}^{\pm 0.008}$	$\mathbf{0.679}^{\pm 0.006}$	$\mathbf{0.644}^{\pm 0.001}$	$\mathbf{0.706}^{\pm 0.002}$

Table 4. **Per-crop semantic correctness evaluation for frame-level Text2Motion generation. Training Set** specifies the dataset used for training, including BABEL, HumanML3D(HML), or the union/intersection of HML and BABEL. **Input** specifies the type of text input. **f**: frame-level text, **s**: sequence-level text. **f+s** demonstrates that combining multi-level conditioning signals can enhance model performance in terms of semantic correspondence. Symbols like ↑ indicates that higher, lower, or values closer to the ground truth (GT) are better, respectively. The evaluation is repeated 10 times, and ± indicates the 95% confidence intervals.

HumanML3D, which contains only squence-level motions. Consequently, FID and Diversity scores from TMR++ offer a more reliable assessment. At the sequence level, both evaluation models yield consistent results. For simplicity and consistency, the main paper presents only FID_TMR++ and Diversity_TMR++.

F. More Applications

Due to space limitations, we only present part of applications in the main paper. Here, we showcase two additional applications that are made possible exclusively by our multi-task model. Similar to UniDiffuser [2], UniMotionnaturally supports applications such as motion variation and text variation. For **motion variation**, given a motion sequence, we first perform the motion-understanding task to generate frame-level text descriptions aligned with the motion. We then use this frame-level text as input for text-to-motion generation, resulting in a new motion that retains similar semantics but with different content. For **text variation**, we reverse the process to produce fine-grained text

Method	Training Set	Input	Per-crop Realism				Per-seq Realism			
			FID ↓	Diversity →	FID_tmr++ ↓	Diversity_tmr++ →	FID ↓	Diversity →	FID_tmr++ ↓	Diversity_tmr++ →
GT	-	-	$0.000^{\pm 0.000}$	$8.823^{\pm 0.067}$	$0.000^{\pm 0.000}$	$1.375^{\pm 0.005}$	$0.000^{\pm 0.000}$	$9.296^{\pm 0.086}$	$0.000^{\pm 0.000}$	$1.391^{\pm 0.003}$
TEACH	BABEL	f	$2.557^{\pm 0.016}$	$7.879^{\pm 0.119}$	$0.155^{\pm 0.001}$	$1.340^{\pm 0.003}$	$3.577^{\pm 0.025}$	$7.605^{\pm 0.066}$	$0.304^{\pm 0.001}$	$1.344^{\pm 0.003}$
DoubleTake	BABEL	f	$2.820^{\pm 0.127}$	$8.248^{\pm 0.102}$	$0.195^{\pm 0.002}$	$1.332^{\pm 0.005}$	$5.619^{\pm 0.268}$	$7.350^{\pm 0.074}$	$0.353^{\pm 0.002}$	$1.337^{\pm 0.004}$
STMC	HML	f	$2.161^{\pm 0.008}$	$9.250^{\pm 0.130}$	$0.156^{\pm 0.000}$	$1.358^{\pm 0.005}$	$1.295^{\pm 0.017}$	$8.955^{\pm 0.102}$	$0.233^{\pm 0.000}$	$1.362^{\pm 0.005}$
FlowMDM	BABEL	f	$0.885^{\pm 0.043}$	$8.476^{\pm 0.086}$	$0.101^{\pm 0.001}$	$1.352^{\pm 0.006}$	$1.028^{\pm 0.060}$	$8.691^{\pm 0.127}$	$0.211^{\pm 0.002}$	$1.375^{\pm 0.005}$
Ours	BABEL	f	$1.206^{\pm 0.079}$	$\underline{9.007}^{\pm 0.141}$	$0.087^{\pm 0.002}$	$1.366^{\pm 0.009}$	$0.791^{\pm 0.091}$	$8.899^{\pm 0.159}$	$0.180^{\pm 0.004}$	$1.374^{\pm 0.002}$
Ours	HML-BABEL	f	$\underline{0.506}^{\pm 0.024}$	$\mathbf{8.979}^{\pm 0.095}$	$\underline{0.071}^{\pm 0.001}$	$\underline{1.372}^{\pm 0.005}$	$\underline{0.401}^{\pm 0.030}$	$\underline{8.956}^{\pm 0.123}$	$\underline{0.150}^{\pm 0.001}$	$\underline{1.378}^{\pm 0.003}$
Ours	HML-BABEL	f + s	$\mathbf{0.487}^{\pm 0.021}$	$9.040^{\pm 0.118}$	$\mathbf{0.066}^{\pm 0.002}$	$\mathbf{1.373}^{\pm 0.009}$	$\mathbf{0.299}^{\pm 0.023}$	$\mathbf{8.978}^{\pm 0.095}$	$\mathbf{0.133}^{\pm 0.004}$	$\mathbf{1.381}^{\pm 0.006}$

Table 5. **Frame-level Text2Motion generation per-crop and per-sequence realism evaluation.** Crop-level realism measures the metrics within each atomic crop, while Seq-level realism measures the fidelity of the overall motion. Symbols ↓, and → indicate that lower, or values closer to the ground truth (GT) are better, respectively.

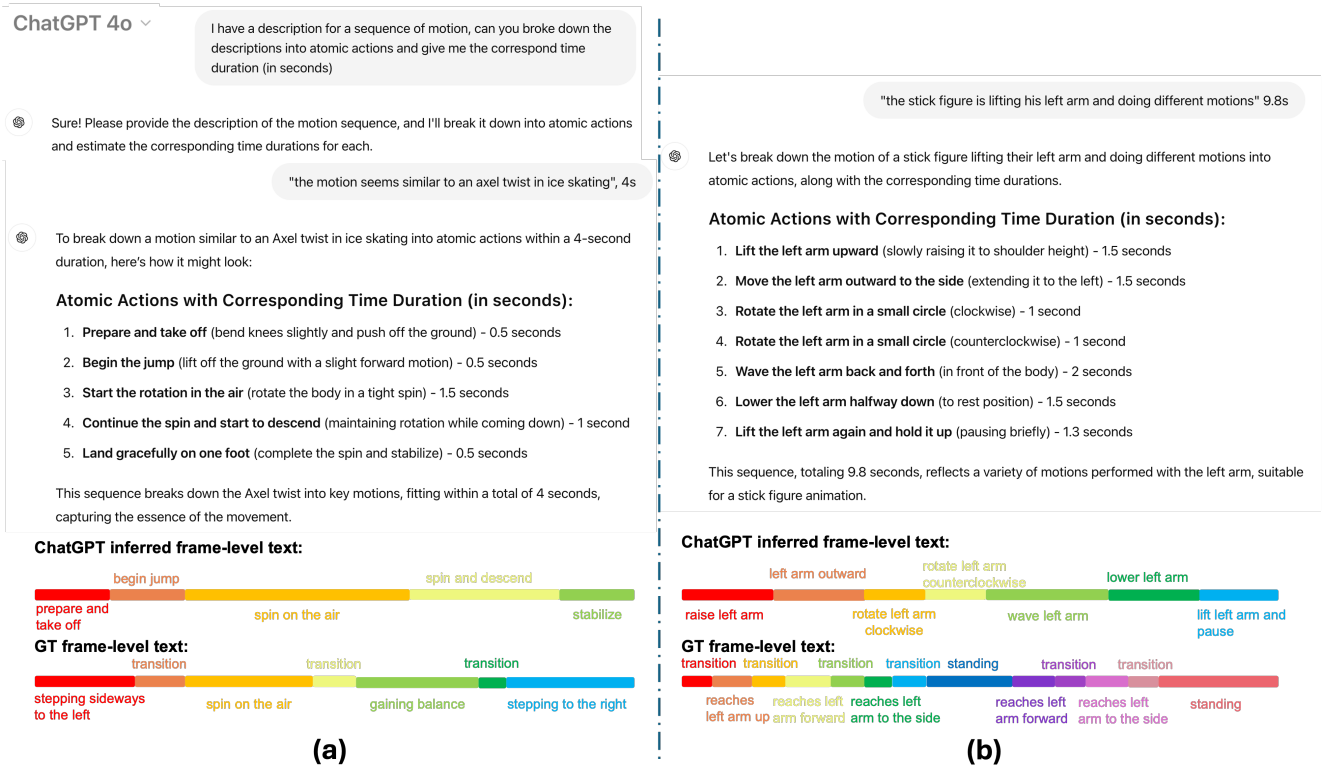

Figure 10. **Fine-grained motion understanding with LLM.** ChatGPT-4o is used to break down the ground-truth global descriptions into atomic motion and durations. However, there is no alignment between text and motion since the model doesn't take the motion as input.

annotation variance. Figure 9 provides examples of both motion and text variation. For animated results, please refer to the attached videos.

G. Motion-to-text Understanding Baselines

To establish baselines for our frame-level motion understanding sub-task, we initially attempted to use a large language model (LLM), ChatGPT, to decompose sequence-level inputs and assess potential outputs. However, due to the LLM's lack of motion awareness, the outputs were unreliable when the sequence-level information was vague or incomplete. Even with detailed sequence-level descriptions, the LLM struggled to generate accurate timestamps due to the absence of motion data. Please refer to Fig. 10 for more details.

We then considered using LLM-based motion models like MotionGPT [17], which can process both motion data and text prompts (to request timestamps and atomic text labels). Despite this, MotionGPT also failed in this task. See Fig. 11 for further information.

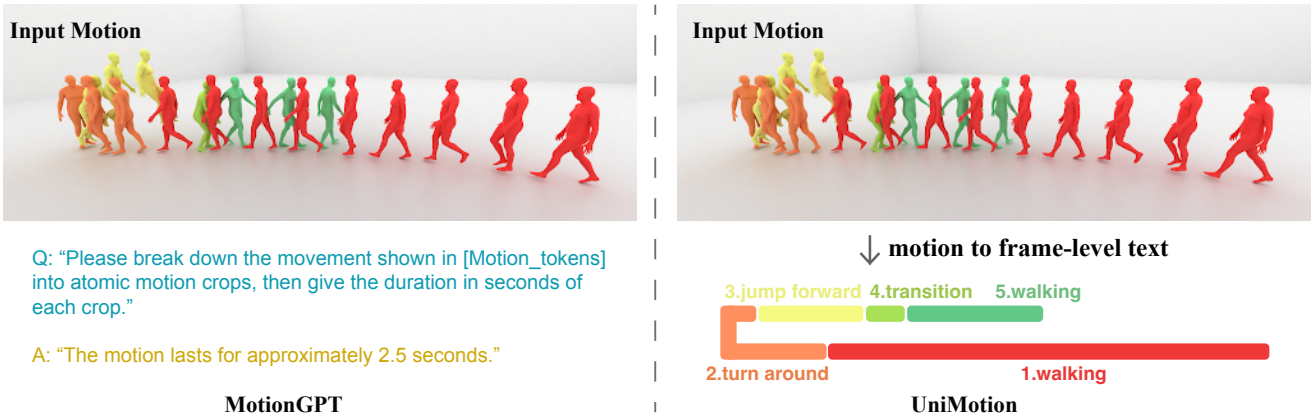

Figure 11. **Motion understanding comparison with MotionGPT [17].** MotionGPT is capable of performing multiple tasks, including motion captioning and question answering. We tasked both MotionGPT (left) and Unimotion (right) with understanding an input motion by breaking it down into motion segments. However, due to MotionGPT's lack of temporal awareness, it was unable to successfully complete this task. Specifically, instead of answering with multiple motion segments, it just predicts an incorrect length for the whole sequence (A: "The motion lasts for approximately 2.5 seconds."). In contrast, our model is the first to understand motion both semantically and temporally.