

# Dual Reciprocal Learning of Language-based Human Motion Understanding and Generation

Chen Liang, Zhicheng Shi, Wenguan Wang, Yi Yang\*

State Key Lab of Brain-Machine Intelligence, Zhejiang University

<https://github.com/leonnop/DRL>

## Abstract

*Language-based human motion understanding focuses on describing human motions using natural language descriptions. Conversely, human motion generation aims to generate human motions from textual inputs. Despite significant progress in both fields, further advancements are hindered by two primary challenges: i) Both tasks rely heavily on vast amounts of paired motion-language data for model training. However, human labeling is costly, making it increasingly unsustainable as model scales increase. ii) Existing models often learn the two tasks in parallel. The strong reciprocity between them has not been fully explored. In response, this work proposes Dual Reciprocal Learning (DRL) for language-based human motion understanding and generation. DRL establishes a symmetric learning framework where both tasks collaboratively evolve in a closed-loop, bootstrapping manner, effectively leveraging the reciprocity between them. In DRL, the tasks serve as evaluators for each other, enabling the generation of informative feedback signals even with easily acquired unpaired, unidirectional motion or language data. Furthermore, to mitigate dataset-specific bias in existing evaluations, we propose a generalized protocol that extends evaluation to a general-domain cross-modal feature space. Experimental results on standard benchmarks demonstrate that DRL achieves remarkable performance boosts over representative baselines in both tasks across evaluation protocols.*

## 1. Introduction

Recent years have witnessed a surge of interest in the convergence of human motion and language modeling, driven by the diverse demands of fields such as robotics [1], game development [2], virtual reality [3, 4], film production [5]. Within this domain, language-based human motion understanding [1, 6–9] and generation [8–16] stand out as pivotal

problems with a long-standing research history in literature. Language-based human motion understanding aims to translate 3D human motions into natural language descriptions or instructions, notably employed for guiding robots or virtual game avatars [1–3]. Conversely, language-based human motion generation focuses on synthesizing desired human motions given the textual guidance. Recent research endeavors have delved deeply into these tasks, centering on sophisticated model architectures [1, 6, 10–14, 16] and training strategies [7, 15]. More recently, a new trend has emerged in scaling up model sizes and capabilities [8, 9], echoing the breakthroughs seen in large foundation models [17–19] to drive further progress in this domain.

While significant advancements have been made in both tasks, two major barriers hinder further progress in the field: ❶ Despite technological strides, both tasks heavily depend on vast amounts of paired motion-language data for model training. Recent progress indicates a growing trend of utilizing more data to train larger models with increased parameters. However, collection of such data is prohibitively expensive and labor-intensive, potentially limiting related research and applications. Besides, ❷ existing studies commonly treat the two tasks as separate problems (Fig. 1 (a)), disregarding the significant interdependence between them. While some initial efforts have aimed to connect the tasks using methods such as one-way alignment [20] (Fig. 1 (b)) or in a unified model design [8, 9] (Fig. 1 (c)), these approaches still tackle the two tasks in parallel, with the motion understanding often being viewed as merely a supplementary component to enhance the learning of motion generation. The strong reciprocity between these two tasks remains largely underexplored in current literature.

To respond, in this work, we propose Dual Reciprocal Learning (DRL) for language-based human motion understanding and generation, denoted as T2M (Text-to-Motion) and M2T (Motion-to-Text). DRL establishes a symmetric learning framework where both tasks collaboratively evolve in a closed-loop, bootstrapping manner, effectively capitalizing on their reciprocal relationship, *cf.*, Fig. 1 (d). Specif-

\*Corresponding author: Yi Yang.

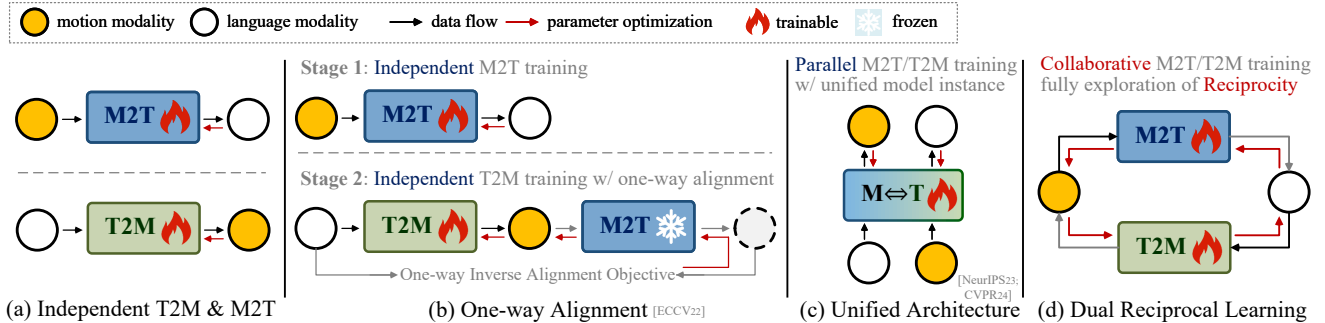


Figure 1. Previous studies commonly treat language-based human motion understanding (M2T, Motion-to-Text) and generation (T2M, Text-to-Motion) as separate problems, *i.e.*, (a). Even though some initial efforts try to connect the two tasks, *i.e.*, (b) and (c), the strong reciprocity between these two tasks remains largely underexplored. In this work, we propose dual reciprocal learning, wherein the two tasks collaboratively evolve in a closed-loop, bootstrapping manner, *i.e.*, (d). Additional details are provided in §1.

ically, we maintain two separate model instances for T2M and M2T. Beginning with either modality (text for T2M or motion for M2T), input data  $x$  is first fed into primal model  $f$  to derive an intermediary output  $f(x)$  in the dual modality, which is then converted back to the primal modality using the dual model  $g$ , *i.e.*,  $\hat{x} = g \circ f(x)$ . In DRL, the dual model  $g$  serves as an evaluator to measure the quality of generated results, and the discrepancy between  $x$  and  $\hat{x}$  is used as feedback signal for model learning. As seen, the training data is not limited to a specific modality or necessarily to be paired inputs, allowing us to effectively learn with both unpaired text and motion data. DRL cycle can be repeated for an arbitrary number of rounds, facilitating continuous enhancement for both model instances. To define discrepancy measurement, DRL involves evaluating semantic similarity in the embedding space for motion modality and assessing the discrete tokenized space for text modality using non-differentiable metrics, *e.g.*, CIDEr [21].

DRL framework has three distinctive features. **First**, our proposed DRL allows unimodal unpaired data to play a similar role to the aligned text-motion pairs, that reduces the dependency on labeled pairwise data during training process, *thereby alleviating the issue one*. **Second**, distinguished from existing efforts (*i.e.*, Fig. 1 (b) and (c)), where M2T model is either fixed or separately trained, thereby only partially and implicitly exploring reciprocal relationship, DRL enhances both models collectively and collaboratively through mutual feedback. Each task serves as the evaluator for the other, whereby the growth in the model’s capabilities corresponds to a proportional increase in the evaluation strengths of both tasks, thereby enforcing further model advancements. Consequently, this iterative process progressively and densely strengthens the reciprocity between the two tasks, *mitigating the issue two*. **Third**, DRL is a principled framework that seamlessly integrates with mainstream T2M and M2T models, which allows DRL to adapt continuously alongside advancements in architecture.

For thorough evaluation, we first review conventional evaluation protocol (*Ori.*) [20], which employs dataset-specific text and motion encoders contrastively trained on each evaluation dataset to measure the feature-space alignment between generated text-motion pairs. While effective within constrained domains, these two specialized encoders introduce inherent evaluation bias: their text-motion alignment favors generations conforming to encoder training set’s kinematic distribution and lexical patterns, thus hindering evaluation on semantic richness and kinematic diversity of novel motions. In response, we propose a generalized evaluation protocol (*Upd.*) that first expands the encoder training corpus to include general-domain data (HumanML3D [20] + Motion-X [22]), and then anchors text-motion alignment to a fixed DistilBERT text encoder [23] to preserve semantic generality while learning motion representations through contrastive objectives analogous to *Ori.*

We evaluate DRL under both protocols by applying DRL over two representative language-based motion understanding and generation baseline models and test it on the gold-standard benchmark, *i.e.*, HumanML3D [20]. Under *Ori.* evaluation, DRL achieves consistent improvements of -0.012 FID and +7.43 CIDEr, confirming its efficacy. While *Upd.* protocol reveals amplified benefits of -0.024 FID and +0.011 RP Top-1, demonstrating its superior generalization beyond dataset-specific artifacts. Extensive experiments show that even in the absence of paired labels, DRL consistently enhances model performance across various architectures for both tasks simultaneously. This not only reaffirms the utility and versatility of DRL but also highlights its potential in real-world applications where unpaired data sources are more readily accessible.

## 2. Related Work

**Human Motion Generation** is generally categorized into unconstrained [24–26] and conditional motion synthesis,

and recent studies mainly focus on the latter. Conditional motion synthesis varies by stimulus type, including coarse action label [27–34], acoustic [35–48] and movement trajectory [49]. In recent years, text-to-motion has been a key focus of research in this field. Recent advances in text-to-motion can be categorized into three mainstreams: (i) VAE-based methods [10, 11] are mainly based on distribution alignment between language and motion latent spaces by applying Kullback-Leibler divergences and contrastive loss, using a decoder to generate motions. (ii) Diffusion-based methods [12–14] leverage denoising diffusion models to synthesize human motion distributions from Gaussian noise guided by text conditions. (iii) Autoregressive methods [15, 16, 50, 51], e.g., T2M-GPT [15], predict motion sequences similarly to language models, conditioning on text inputs and previous motion tokens. The latest works, e.g., MotionGPT [8] and AvatarGPT [9], exploit large language models to predict motion token sequences.

Despite promising results, existing approaches primarily focus on designing sophisticated model architectures for SOTA performance on generation quality and text-motion alignment. For instance, MoMask [50] adopts a hierarchical quantization scheme with two well-designed transformers, while BAMB [51] uses a masked transformer with hybrid attention masking to predict randomly masked tokens. However, existing methods still rely on text-motion pairs and are potentially constrained by the limited availability of paired text-motion data. Meanwhile, abundant unpaired motion data remains largely underexplored.

**Human Motion Understanding.** Compared to motion generation, studies on human motion understanding remain comparatively scarce. InfoGCN [52] and MotionBERT [53] identify the pre-defined action category from motion inputs. Other approaches leverage natural language, with early works [1, 6] using RNN-based models to generate coarse textual descriptions. Later in TM2T [7], transformer is explored operating on discretized motion tokens.

Recent works [8, 9] have introduced unified frameworks for text-to-motion and motion-to-text using pretrained large language models. Among them, AvatarGPT [9] leverages large-scale LLM model with additional video-text data and achieves SOTA performance on semantic richness and expression coherence. However, these unified methods focus more on text-to-motion, only viewing the motion-to-text model as a by-product during training process. In contrast, DRL treats motion-to-text and text-to-motion equally to leverage their intrinsic connections and form a closed loop for end-to-end joint training to bootstrap each other.

**Dual Learning** [54] originates from neural machine translation (NMT). The core idea involves establishing a two-agent communication game. Each agent, assigned an individual task, is required to map an  $x$  from the primal domain to its dual domain  $y$  and then recover the original  $x$  through

the reverse mapping in dual task, thus producing a feedback signal without parallel data [55, 56]. Dual learning has moderately addressed the dependence of deep learning methods on data annotation and has been successfully applied to a variety of cross-modality generation tasks, such as visual question answering [57, 58], visual acoustic matching [59], and vision language navigation [60]. Noticing that motion generation and understanding have a reciprocal relationship, which is analogous to that of bidirectional language translation, we propose a dual-task bootstrapping framework in which we could emphasize and explore the structural symmetry of two correlated tasks and capitalize on one-sided data samples for yielding informative feedback signals to boost each other.

### 3. Methodology

We propose a dual learning framework, namely DRL, to leverage feedback signals from symmetrical tasks to promote model learning and better exploit unpaired data. We first give a brief preliminary on T2M and M2T tasks (§3.1), then illustrate the general formulation of DRL (§3.2). DRL is a principled framework that is not tied to specific model design. In §3.3, we present one instantiation of DRL, adopting representative architectures drawn from mainstream T2M and M2T models. Finally, in §3.4, we provide the implementation details of the proposed pipeline.

#### 3.1. Preliminary: T2M and M2T Tasks

DRL involves two tasks: the primal task motion generation [10] that employs the T2M model  $f_\theta$  to convert textual inputs into the desired motion sequences. In the dual task, motion understanding [1], M2T model  $g_\phi$  translates motion sequences back to corresponding textual descriptions.

**T2M.** Considering paired text-motion data distribution, denoted as  $\mathcal{D}^L = \{(t_n, v_n)\}_{n=1}^N$ , the goal of T2M is to convert the textual description  $t_n$  to corresponding motion sequence  $v_n$ , i.e., to estimate the conditional distribution  $f_\theta(v_n|t_n)$ . The network parameters  $\theta$  are optimized with  $\theta^* = \operatorname{argmax}_\theta \sum_{(t,v) \in \mathcal{D}} \log P(v|t, \theta) = \operatorname{argmin}_\theta \mathcal{L}_{\text{T2M}}(t)$ .

**M2T.** Contrarily, M2T model  $g_\phi$  translates motion sequences into text descriptions  $g_\phi(t_n|v_n)$ , that seeks optimal parameters  $\phi^*$  with  $\phi^* = \operatorname{argmax}_\phi \sum_{(t,v) \in \mathcal{D}} \log P(t|v, \phi) = \operatorname{argmin}_\phi \mathcal{L}_{\text{M2T}}(v)$ . As seen, conventionally, both tasks are independently trained on paired text-motion inputs, ignoring the reciprocity between the two tasks while not being able to utilize the unpaired, one-way data.

#### 3.2. Dual Reciprocal Learning Framework

Distinct from conventional T2M/M2T models, DRL learns T2M and M2T tasks jointly. The goal is to simultaneously achieve and progressively enhance both motion understanding and motion generation by utilizing paired and unidirectional non-paired data. To achieve such goal, we exploit the

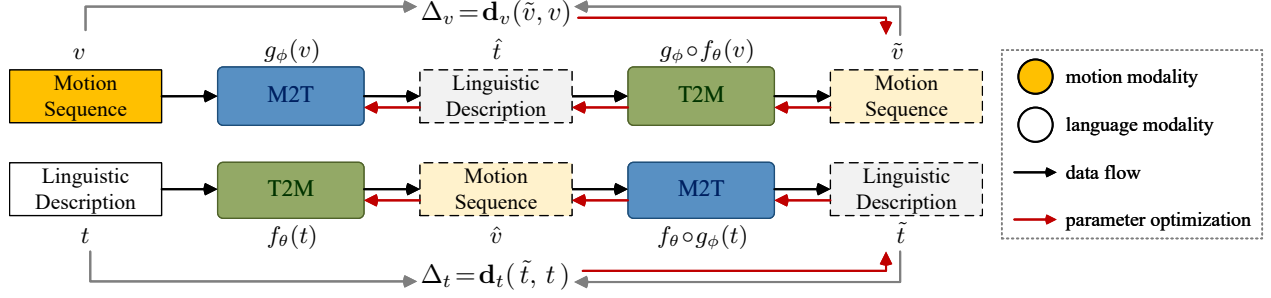


Figure 2. Framework overview of DRL. Starting from either task, data samples are first forwarded to another domain, then reversed back to the original. Within this iterative process, two models act as mutual evaluators, providing feedback signals that facilitate collaborative learning even with unpaired data from single modality.  $\mathbf{d}(\cdot, \cdot)$  is the distance measurement. For a comprehensive elaboration, refer to §3.2.

reciprocity between these two tasks, where the input-output spaces of T2M and M2T demonstrate a strong correlation and can alternatively serve as the evaluator for each other.

Starting from either task, we initially transform a unidirectional data sample forward to another domain, then reverse the transfer to original domain. Dual model is utilized to assess the quality of intermediate outputs produced by primal model and transmits an error signal back to the primal, and vice versa. By evaluating results of this two-step transfer, we can assess the quality of both models and refine them correspondingly. This iterative process can be repeated multiple times until both models converge.

More formally, for the primal process starting from text modality, we first sample random text example  $t$  and  $\hat{v} = f_\theta(t)$  is the mid-transition output. Then, dual model  $g_\phi$  translates  $\hat{v}$  to  $\tilde{t}$  by mapping  $g_\phi(\hat{v})$ . The  $\tilde{t}$  is expected to be consistent with  $t$  in semantic meaning, *i.e.*, achieving a small cycle-consistency error  $\Delta_t = \mathbf{d}_t(\tilde{t}, t)$ , where  $\mathbf{d}(\cdot, \cdot)$  is the distance measurement. Conversely, starting from the dual task, we have  $\tilde{v} = f_\theta \circ g_\phi(v)$  and  $\tilde{v}$  should be akin to  $v$  in motion fidelity. Likewise,  $\Delta_v$  can be employed to evaluate the discrepancies between  $v$  and  $\tilde{v}$ . Finally, the errors  $\Delta_t$  and  $\Delta_v$  can be specified as two reconstruction losses, which are minimized for model training.

As discussed,  $t$  and  $v$  are not necessarily aligned. This allows exploration of unimodal unpaired data, addresses the over-reliance on labeled data pairs during the training process. More importantly, DRL is general in its formulation. Instantiations of  $f_\theta$ ,  $g_\phi$ , and  $\Delta_{v/t}$  are not restricted to specific model or loss design. In the next section, we demonstrate specific application of DRL on existing models.

### 3.3. Framework Instantiation

Given the labeled collection  $\mathcal{D}^L = \{(t_n^L, v_n^L)\}_{n=1}^N$ , unpaired unimodal text data  $\mathcal{U}_T = \{(t_m^U)\}_{m=1}^M$  and unimodal motion data  $\mathcal{U}_V = \{(v_k^U)\}_{k=1}^K$ , consisting of  $N$  aligned text-motion pairs,  $M$  and  $K$  unimodal data samples, respectively. T2M model  $f_\theta$ , M2T model  $g_\phi$ , discrepancy measurement  $\Delta_{t/v}$

and framework optimization are instantiated as follows.

**T2M Model.** Latest methods often follow a quantization-prediction pipeline: Motions are first quantized into discrete motion embeddings using a discrete VAE (*e.g.*, VQ-VAE [61]). Then, a sequence-to-sequence network (*e.g.*, autoregressive or diffusion-based) predicts the quantized motion, conditioned on text feature extracted by large-scale pretrained text encoder (*e.g.*, CLIP [62]). This work adopts a representative architecture as follows.

Define a motion tokenizer composed of an encoder  $\mathcal{F}_{\text{Enc}}$ , a decoder  $\mathcal{F}_{\text{Gen}}$ , and a learned discrete codebook  $Z$ . For a motion sample  $v$ , drawn from both labeled and unlabeled motion sets (*i.e.*,  $v \in \{v_n^L\}_{n=1}^N \cup \{v_k^U\}_{k=1}^K$ ), it is fed into  $\mathcal{F}_{\text{Enc}}$ , which outputs a latent feature sequence. Each element in this sequence is quantized by finding its nearest neighbor vector in  $Z$ , yielding a corresponding sequence of indices  $s = \text{Index}(Z, \mathcal{F}_{\text{Enc}}(v))$ . The indices  $s$  represent the discrete motion codes. The original motion can be approximately reconstructed as  $v' \approx \mathcal{F}_{\text{Gen}}(Z[s])$ , where  $Z[s]$  maps each index in  $s$  back to its corresponding codebook vector.

Next, autoregressive next-index prediction network  $\mathcal{F}_\theta$  is defined for T2M generation. Given the previous  $i-1$  indices  $s_{<i}$  and text condition  $c = \mathcal{F}_{\text{CLIP}}(t)$  encoded from text description  $t$ ,  $\mathcal{F}_\theta$  predicts the distribution of possible next index  $s_i$ :  $\mathcal{F}_\theta(s_i | c, s_{<i})$ . The text  $t$  is the corresponding description for  $v$ , which comes either from the ground-truth label  $t_n^L$  paired with  $v_n^L$ , or is an intermediate output generated for unlabeled motion  $v_k^U$  by the dual M2T model  $g_\phi$ , *i.e.*,  $t \in \{t_n^L\}_{n=1}^N \cup \{g_\phi(v_k^U)\}_{k=1}^K$ . During T2M training, only parameters of autoregressive network  $\mathcal{F}_\theta$  are trained.

**M2T Model.** Similarly, M2T model  $\mathcal{F}_\phi$  follows an autoregressive next-token prediction paradigm, conditioned on the encoded discrete motion token sequence  $s$ :  $\mathcal{F}_\phi(t_i | s, t_{<i})$ .

**Discrepancy Measurement  $\Delta_{t/v}$ .**  $\Delta_{t/v}$  measures discrepancies between the estimated results and ground-truths. Depending on their formulation, they can be either differentiable or not. For  $\Delta_v$ , we compute the cross-entropy between generated motion tokens  $\tilde{s}$  and original  $s$ :  $\Delta_v =$

$\mathcal{L}_{CE}(s, \tilde{s})$ . Since  $\tilde{s}$  is generated by the differentiable prediction network  $\mathcal{F}_\theta$  (and the motion tokenizer  $\mathcal{F}_{\text{Enc}}, \mathcal{F}_{\text{Gen}}, Z$  is fixed),  $\Delta_v$  is differentiable, enabling end-to-end training via backpropagation. For  $\Delta_t$ , we define it as sum of normalized CIDEr [21] and BertScore [63] between the generated text  $\tilde{t}$  and the reference. Due to its reliance on non-differentiable semantic evaluation metrics,  $\Delta_t$  is non-differentiable. In this case, we use reinforcement learning, *i.e.*, policy gradient optimization, where  $\Delta_t$  serves as the reward signal.

**Two-stage Optimization.** To bootstrap DRL, at stage one, the framework is optimized using labeled text-motion pairs  $(t^L, v^L) \in \mathcal{D}^L$  and unlabeled motions  $v^U \in \mathcal{U}_V$ :

$$\mathcal{L}_{S1} = \mathcal{L}_{\text{labeled}} + \lambda_V \mathcal{L}_V. \quad (1)$$

Here,  $\mathcal{L}_{\text{labeled}}$  supervises paired data:

$$\begin{aligned} \mathcal{L}_{\text{labeled}} &= \frac{1}{N} \sum_{(t^L, v^L) \in \mathcal{D}} (\mathcal{L}_{\text{T2M}}(t^L) + \mathcal{L}_{\text{M2T}}(v^L)), \\ \mathcal{L}_{\text{T2M}} &= - \sum_{i=0}^{|s^L|} \log \mathcal{F}_\theta(s_i^L | c^L, s_{<i}^L), \\ \mathcal{L}_{\text{M2T}} &= - \sum_{i=0}^{|t^L|} \log \mathcal{F}_\phi(t_i^L | s^L, t_{<i}^L). \end{aligned} \quad (2)$$

$\mathcal{L}_V$  leverages unlabeled motions:

$$\mathcal{L}_V = \frac{1}{K} \sum_{v^U \in \mathcal{U}_V} \Delta_{v^U} = \frac{1}{K} \sum_{v^U \in \mathcal{U}_V} \mathcal{L}_{CE}(s^U, \tilde{s}^U). \quad (3)$$

$\mathcal{L}_{\text{labeled}}$  and  $\mathcal{L}_V$  are balanced by weighting parameter  $\lambda_V$ .

At stage two, DRL utilizes *only* unlabeled texts  $t^U \in \mathcal{U}_T$  to generate feedback and learns via RL. We define a reward measuring text semantic similarity:

$$\mathcal{R}(\tilde{t}^U) = \text{CIDEr}(t^U, \tilde{t}^U) + \text{Bert-S}(t^U, \tilde{t}^U), \quad (4)$$

which is maximized via policy gradient:

$$\mathcal{L}_{S2} = \mathcal{L}_T = -\mathbb{E}_{\tilde{t}^U \sim (f_\theta, g_\phi)} [\mathcal{R}(\tilde{t}^U)] \quad (5)$$

utilizing SCST [64] with PPO [65]. This two-stage design enhances training stability and explores text-motion reciprocity with unpaired data throughout optimization.

### 3.4. Implementation Details

**Network Architecture.** We adopt VQ-VAE [61] as discrete VAE and CLIP [62] for text condition extraction, and employ the same architecture as [15] for  $\mathcal{F}_{\text{Enc}}, \mathcal{F}_{\text{Gen}}, Z$  and  $\mathcal{F}_{\text{CLIP}}$ . T2M/M2T autoregressive networks  $\mathcal{F}_\theta/\mathcal{F}_\phi$  are implemented as standard Transformer [66] (See §4.1).

**Training.** DRL facilitates cyclic propagation of gradients between T2M and M2T models. While this gradient flow could, in theory, iterate indefinitely, we truncate the back-propagation after the first cycle for computational efficiency and implementation stability. We pretrain the motion VQ-VAE following the same training process to [15] and keep it fixed in DRL learning.  $\lambda_V$  is set to 1.0.

**Reinforcement Learning.** RL setups largely follow [67]. We treat T2M-M2T model pair as policy model and add a KL penalty weighted by  $\beta$  to constrain the policy’s output from deviating too far from that of a frozen pretrained M2T

model obtained in stage one. We disable dropout during policy training [67, 68], and use a batch size of 128, learning rate of  $1 \times 10^{-7}$ ,  $\beta = 0.02$ ,  $\gamma = 1$ , and four PPO epochs per batch. Other hyperparameters are set to common default.

**Inference.** Following conventions [8, 15], we adopt probabilistic sampling to generate non-deterministic outputs for both T2M and M2T. Notably, DRL jointly trains T2M and M2T instances with the same FLOPs as baselines, which train each single model in parallel. DRL deploys them independently in inference without adding extra parameters.

## 4. Experiments

### 4.1. Experimental Setup

**Dataset.** We use the gold-standard text-motion dataset, HumanML3D [20]. It provides 14,616 motions and 44,970 texts composed by 5,371 distinct words. Each motion is annotated with at least 3 descriptions and is scaled to 20 FPS.

**Unimodal Data.** We establish unimodal data ( $\mathcal{U}_V$  for motion and  $\mathcal{U}_T$  for text) with Motion-X [22] dataset. Motion-X comprises 81.1K high-quality, diverse motion clips (including locomotion, object interaction, dancing, and martial arts) paired with 81.1K semi-automatically generated sequence-level text descriptions, of which 13,581 are refined for fine-grained semantics. We randomly select 12K motion clips as  $\mathcal{U}_V$  and the full set of 13,581 refined texts as  $\mathcal{U}_T$  to balance performance and training cost. Motion samples use the SMPL-X format [69], which includes body, finger, and facial keypoints; to align with HumanML3D’s representation, we retain only body keypoints.

**Baseline Setting.** We set representative baselines (Fig. 1):

- **One-way-align Methods** [7, 15] perform T2M or M2T separately. We adapt TM2T [7], replacing its T2M and fixed M2T model [70] with same Transformer model from T2M-GPT [15]. This baseline is denoted as TM2T\*.
- **Unified Methods** [8, 9] use a single model for both tasks. Original MotionGPT [8] handles additional tasks like motion completion and prediction. We adapt its T5 [71] architecture solely for T2M-M2T, termed MotionGPT\*.

**Training Details.** We adopt the pre-processing, optimizer, and batch size configurations from [15] and [8] for training TM2T\* and MotionGPT\*, respectively. Both baselines are trained for 800K iterations with official learning rate schedule for each. M2T decoding uses a maximum token length of 128. All experiments are conducted on four NVIDIA GeForce RTX 4090 GPUs with PyTorch. More details are available in our github repository<sup>1</sup>.

**Evaluation Protocol.** Standard protocol evaluates T2M/M2T outputs in a text-motion aligned embedding space for fidelity, diversity, and cross-modal alignment. Conventional works (*Ori.*) [20] train separate text and motion encoders

<sup>1</sup><https://github.com/leonnop/DRL>

Method	R Precision $\uparrow$			FID $\downarrow$	MM Dist $\downarrow$	Diversity $\uparrow$	MModality $\uparrow$
	Top 1	Top2	Top3				
<b>Real motions</b>	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 [7] [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 [12] [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 [14] [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 [15] [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 [16] [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
MoMask [50] [CVPR24]	0.521 $\pm$ .002	0.713 $\pm$ .002	0.807 $\pm$ .002	0.045 $\pm$ .002	2.958 $\pm$ .008	-	1.241 $\pm$ .040
BAMM [51] [ECCV24]	0.525 $\pm$ .002	0.720 $\pm$ .003	0.814 $\pm$ .003	0.055 $\pm$ .002	2.919 $\pm$ .008	9.717 $\pm$ .089	1.687 $\pm$ .051
MotionGPT [8] [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 [9] [CVPR24]	0.510 $\pm$ .005	0.702 $\pm$ .005	0.796 $\pm$ .003	0.168 $\pm$ .008	-	9.624 $\pm$ .055	-
TM2T*-T2M	0.463 $\pm$ .004	0.638 $\pm$ .002	0.742 $\pm$ .005	0.092 $\pm$ .002	3.291 $\pm$ .010	9.453 $\pm$ .034	1.882 $\pm$ .048
TM2T*-T2M + DRL	<b>0.468<math>\pm</math>.003</b>	0.634 $\pm$ .002	0.735 $\pm$ .004	<b>0.080<math>\pm</math>.004</b>	3.361 $\pm$ .016	9.398 $\pm$ .042	<b>2.095<math>\pm</math>.037</b>
MotionGPT*	0.487 $\pm$ .004	0.677 $\pm$ .005	0.776 $\pm$ .004	0.363 $\pm$ .005	3.112 $\pm$ .009	9.430 $\pm$ .045	2.280 $\pm$ .040
MotionGPT* + DRL	<b>0.490<math>\pm</math>.003</b>	<b>0.679<math>\pm</math>.004</b>	<b>0.777<math>\pm</math>.003</b>	<b>0.327<math>\pm</math>.003</b>	3.125 $\pm$ .008	9.396 $\pm$ .037	<b>2.425<math>\pm</math>.029</b>

Table 1. Quantitative comparison of motion generation (§4.2) on HumanML3D test under *Ori.* protocol (§4.1). Metrics computed using encoders from [20], which is biased to specific HumanML3D data distribution.  $\pm$ : 95% confidence interval. **Blue**: DRL improvements. (\*) denotes our re-trained baselines (§4.1); other results directly borrowed from respective papers.

Method	R Precision $\uparrow$			FID $\downarrow$	MM Dist $\downarrow$	Diversity $\uparrow$	MModality $\uparrow$
	Top 1	Top2	Top3				
<b>Real motions</b>	0.386 $\pm$ .002	0.568 $\pm$ .003	0.687 $\pm$ .002	0.001 $\pm$ .000	3.661 $\pm$ .010	8.026 $\pm$ .023	-
TM2T*-T2M	0.370 $\pm$ .002	0.552 $\pm$ .003	0.666 $\pm$ .003	0.063 $\pm$ .002	3.651 $\pm$ .034	8.101 $\pm$ .048	2.091 $\pm$ .035
TM2T*-T2M + DRL	<b>0.380<math>\pm</math>.003</b>	<b>0.557<math>\pm</math>.002</b>	0.643 $\pm$ .004	<b>0.039<math>\pm</math>.003</b>	3.736 $\pm$ .027	<b>8.156<math>\pm</math>.033</b>	<b>2.238<math>\pm</math>.029</b>
MotionGPT*	0.346 $\pm$ .002	0.521 $\pm$ .004	0.631 $\pm$ .004	0.231 $\pm$ .002	3.833 $\pm$ .010	8.021 $\pm$ .045	2.436 $\pm$ .034
MotionGPT* + DRL	<b>0.364<math>\pm</math>.003</b>	<b>0.552<math>\pm</math>.004</b>	<b>0.664<math>\pm</math>.005</b>	<b>0.190<math>\pm</math>.002</b>	<b>3.701<math>\pm</math>.029</b>	<b>8.069<math>\pm</math>.035</b>	<b>2.610<math>\pm</math>.035</b>

Table 2. Quantitative comparison of motion generation (§4.2) on HumanML3D test set using *Upd.* protocol (§4.1). Metrics computed with re-trained encoders on generalized distributions. Decline in RP-3 and MM Dist may stem from diversified outputs, reducing exact matches.

per benchmark using only its training data, limiting the evaluation’s generalizability and binding it to dataset-specific biases. To address this, we propose *Upd.* setup that establishes a unified embedding space:

- **Specialized Domain (*Ori.*):** Use dataset-specific, separately pre-trained encoders per benchmark.
- **General Domain (*Upd.*):** Create a unified, generalized embedding space by: (i) Employing a fixed, general-purpose DistilBERT text encoder [23] instead of specialized ones; (ii) Expanding the motion encoder training data to include both HumanML3D and Motion-X, enabling better alignment between diverse motions and texts. This explicitly tackles the *Ori.* setup’s lack of generalization beyond individual benchmarks.

## 4.2. Performance on Motion Generation

**Metrics.** Following conventions [7], we evaluate: (i) Motion quality via Fréchet Inception Distance (FID) [72] between generated and real motions. (ii) Motion diversity using Diversity (DIV), *i.e.*, feature-space variance, and MultiModality (MM), *i.e.*, output variation under identical text prompts. (iii) Text-motion alignment with motion-retrieval

precision (R Precision) [7], *i.e.*, Top-k retrieval accuracy, and Multi-modal Distance (MM Dist). All metrics derive from text/motion encoders and are protocol-dependent.

**Quantitative Results.** We summarize motion generation results in Tables 1-2, highlighting four key observations for DRL-enhanced models: (i) Though affected by the biased *Ori.* protocol, DRL achieves impressive improvements in FID, confirming enhanced generation fidelity. (ii) Under *Upd.* protocol, DRL yields higher R-Precision, reflecting stronger text-motion alignment. (iii) Under both protocols, DRL effectively boosts the overall and class-wise diversity. (iv) A decline in RP Top3 and MM Dist is observed. This may stem from diversified outputs reducing exact matches.

**Qualitative Results.** As in Fig. 3 (row 1), TM2T\*-T2M + DRL generates an exact throwing motion whereas TM2T\*-T2M fails, indicating DRL’s role in enhancing semantic perception and motion diversity. In row 2, motion generated by baseline lacks completeness and precision. It shows simultaneous leg movements and misses a back-to-left motion. In contrast, TM2T\*-T2M + DRL produces a fully aligned motion, demonstrating DRL’s benefit for fine-grained semantic understanding and motion generation.

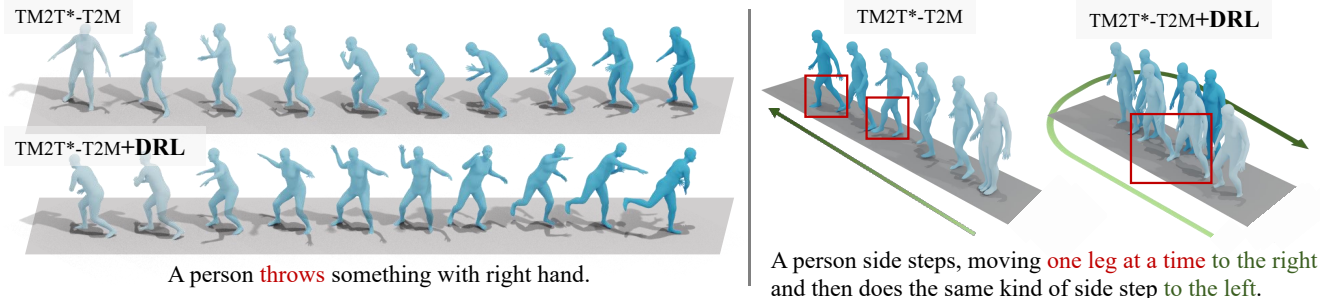


Figure 3. Qualitative comparison of TM2T\* and TM2T\*+DRL on motion generation (§4.2).

Method	R Precision $\uparrow$			MM Dist $\downarrow$	BLEU@1 $\uparrow$	BLEU@4 $\uparrow$	Rouge $\uparrow$	CIDEr $\uparrow$	BertScore $\uparrow$
	Top 1	Top2	Top3						
<b>Real Desc.</b>	0.523	0.725	0.828	2.901	-	-	-	-	-
$\dagger$ TM2T [7] <sub>[ECCV22]</sub>	0.516	0.720	0.823	2.935	48.90	7.00	38.10	16.80	32.20
$\ddagger$ MotionGPT [8] <sub>[NeurIPS23]</sub>	0.543	-	0.827	2.821	48.20	12.47	37.40	29.20	32.40
$\ddagger$ AvatarGPT [9] <sub>[CVPR24]</sub>	-	-	-	-	49.28	12.70	40.44	32.65	53.58
TM2T*-M2T	0.288	0.449	0.547	4.790	23.65	4.42	27.94	43.47	18.03
TM2T*-M2T + DRL	<b>0.303</b>	<b>0.472</b>	<b>0.569</b>	<b>4.622</b>	<b>28.62</b>	<b>5.80</b>	27.68	<b>50.90</b>	<b>19.88</b>
MotionGPT*	0.504	0.687	0.780	3.065	43.44	6.01	34.38	5.68	23.28
MotionGPT* + DRL	<b>0.511</b>	<b>0.697</b>	<b>0.788</b>	<b>3.059</b>	<b>45.70</b>	<b>7.52</b>	<b>35.93</b>	<b>13.26</b>	<b>26.01</b>

Table 3. Quantitative comparison of motion understanding (§4.3) on HumanML3D test (*Ori.* protocol, §4.1).  $\dagger$ : TM2T uses an expanded VQVAE codebook in size and dimensions;  $\ddagger$ : Models trained with enhanced scale/strategies/data augmentation. Results of state-of-the-art methods, sourced from [8] and [9] papers, were evaluated using NLG-Eval [73], while we employ NLG-Metricverse [74] following official implementation of [8]. This evaluation setup discrepancy may lead to inconsistencies when comparing their results with ours.

Method	R Precision $\uparrow$			MM Dist $\downarrow$
	Top 1	Top2	Top3	
<b>Real Desc.</b>	0.384	0.572	0.685	3.591
TM2T*-M2T	0.199	0.324	0.420	5.145
TM2T*-M2T + DRL	<b>0.210</b>	<b>0.347</b>	<b>0.446</b>	<b>5.076</b>
MotionGPT*	0.352	0.538	0.661	3.580
MotionGPT* + DRL	<b>0.363</b>	<b>0.552</b>	<b>0.667</b>	<b>3.565</b>

Table 4. Quantitative comparison of motion understanding (§4.3) on HumanML3D test set under *Upd.* protocol (§4.1).

### 4.3. Performance on Motion Understanding

**Metrics.** For motion-to-text evaluation, we assess two aspects: (i) Text-motion alignment via R-Precision and Multimodal Distance (MM Dist), evaluating M2T at embedding level; (ii) Natural language quality using linguistic metrics, *i.e.*, BLEU [75], ROUGE [76], CIDEr [21] and BertScore [63], computed at the text token level with NLG-Metricverse [74]. Alignment metrics depend on the evaluation protocol, while linguistic metrics are protocol-agnostic.

**Quantitative Results.** Table 3 and 4 compare motion understanding performance under different protocols. DRL-enhanced TM2T\*-M2T surpasses its non-DRL version across four linguistic metrics, most notably in CIDEr (50.90

vs 43.47), demonstrating significant gains in semantic accuracy. Under the *Upd.* protocol, consistent improvements in motion-text alignment metrics confirm DRL’s capability to strengthen cross-modal alignment.

**Qualitative Results.** Fig. 4 (left) shows TM2T-M2T\* + DRL correctly identifies subtle picking-up gestures with detailed descriptions, while TM2T\*-M2T outputs irrelevant content. On the right, the baseline merely tracks trajectory, whereas TM2T\*-M2T + DRL captures full-body semantics to correctly recognize dancing with precise and concise descriptions. These contrasts demonstrate DRL’s role in enhancing fine-grained motion perception, holistic semantic understanding, rich and concise semantic expression.

### 4.4. Diagnostic Experiments

For in-depth analysis, we perform a set of ablative studies with TM2T\*-T2M and TM2T\*-M2T under *Upd.* protocol.

**Dual Learning.** We first validate the significance of both unpaired data and reciprocal relations within our DRL framework. Table 5a (row 1) reports the result of a baseline model trained solely on paired text-motion data. Next, we implement our dual-learning bootstrapping process using only the paired data, *i.e.*, training two models with  $\mathcal{L}_V$  and  $\mathcal{L}_T$  exclusively on text-motion pairs, without incorporating any unpaired data. This yields notable improve-

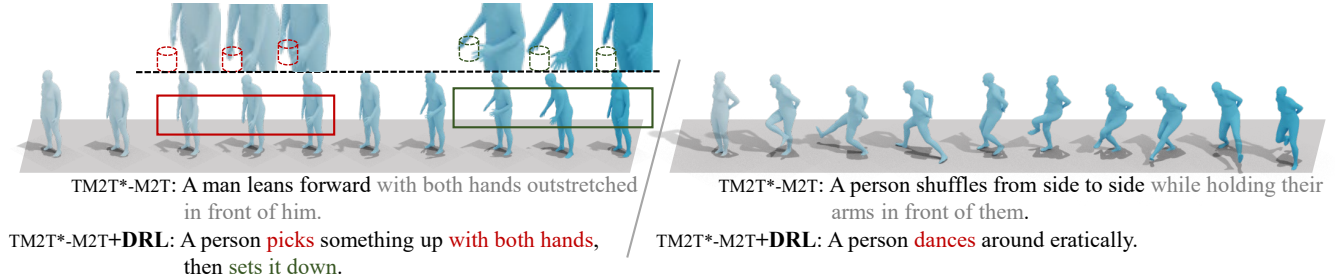


Figure 4. Qualitative comparison of TM2T\*-M2T to TM2T\*-M2T + DRL on motion understanding (§4.3). Gray indicates irrelevant texts.

(a) Dual Learning						(b) Unpaired Data Size					
Method	Text-to-Motion			Motion-to-Text		Data Size	Text-to-Motion			Motion-to-Text	
	FID ↓	R TOP1 ↑	MModality ↑	BLEU@4 ↑	CIDEr ↑		$\mathcal{U}_V$	FID ↓	R TOP1 ↑	MModality ↑	BLEU@4 ↑
Baseline ( $w$ only $\mathcal{D}^L$ )	0.063	0.370	2.091	4.42	43.47	3k	0.056	0.367	2.106	4.75	45.02
DRL <i>w/o</i> additional data	0.048	0.376	2.156	5.11	47.28	5k	0.048	0.371	1.171	5.32	47.45
DRL $w \mathcal{U}_V$	0.043	0.377	2.192	5.45	48.07	9k	0.039	0.380	2.238	5.80	50.90
DRL $w \mathcal{U}_V + \mathcal{U}_T$	0.039	0.380	2.238	5.80	50.90	12k	0.031	0.388	2.265	5.97	52.23

(c) Paired Data Reliance						(d) Training Strategy						
Data Ratio	Text-to-Motion			Motion-to-Text		Training Strategy		Text-to-Motion			Motion-to-Text	
	FID ↓	R TOP1 ↑	MModality ↑	BLEU@4 ↑	CIDEr ↑	Stage One	Stage Two	FID ↓	R TOP1 ↑	MModality ↑	BLEU@4 ↑	CIDEr ↑
$\mathcal{D}^L$	0.063	0.370	2.091	4.42	43.47	$\mathcal{L} + \mathcal{L}_V + \mathcal{L}_T$		0.134	0.352	2.677	4.74	39.52
$1/2\mathcal{D}^L + \mathcal{U}_T + \mathcal{U}_V$	0.064	0.367	1.977	4.36	38.70	$\mathcal{L}$	$\mathcal{L}_V + \mathcal{L}_T$	0.060	0.371	2.107	4.89	45.08
$1/4\mathcal{D}^L + \mathcal{U}_T + \mathcal{U}_V$	0.071	0.361	1.970	4.30	30.16	$\mathcal{L} + \mathcal{L}_V$	$\mathcal{L} + \mathcal{L}_V + \mathcal{L}_T$	0.051	0.377	2.179	5.22	47.95
$1/8\mathcal{D}^L + \mathcal{U}_T + \mathcal{U}_V$	0.105	0.349	1.853	4.26	22.75	$\mathcal{L} + \mathcal{L}_V$	$\mathcal{L}_T$	0.039	0.380	2.238	5.80	50.90

Table 5. A series of ablation studies on HumanML3D test set [20]. All the experiments are conducted under *Upd.* evaluation protocol.

ment (row 2), suggesting that the structural symmetry between primal and dual tasks alone provides valuable training signals. Further incorporating unpaired motion data during dual learning yields additional gains (row 3). Finally, leveraging both unpaired text and motion data within DRL achieves the greatest improvement (row 4), demonstrating the full efficacy of the DRL framework.

**Unpaired Data Size.** Then, we investigate the impact of unpaired motion data volume within DRL. Table 5b demonstrates that increasing unpaired motion data consistently boosts performance, evidencing DRL’s scalability. To balance performance and computational resources, we incorporate 9k motion samples in DRL by default.

**Paired Data Reliance.** Furthermore, we analyze DRL’s reliance on paired data. Using a fixed amount of unpaired data and progressively decreasing amounts of paired data (Table 5c), we observe that the performance of the DRL-enhanced TM2T\*-T2M model remains comparable to the baseline model’s even when paired data is reduced to one-fourth of the original amount. This demonstrates that DRL reduces the dependency on paired data.

**Training Strategy.** Finally, we study the influence of different training strategies in Table 5d.  $\mathcal{L}$  denotes  $\mathcal{L}_{\text{labeled}}$ . We investigate four alternatives: (i) Initialize and maintain  $\mathcal{L} + \mathcal{L}_V + \mathcal{L}_T$  throughout training; (ii) Stage One:  $\mathcal{L}$ ;

Stage Two:  $\mathcal{L}_V + \mathcal{L}_T$ ; (iii) Stage One:  $\mathcal{L} + \mathcal{L}_V$ ; Stage Two:  $\mathcal{L} + \mathcal{L}_V + \mathcal{L}_T$ ; (iv) The DRL strategy proposed in §3.3. In practice, we find our proposed strategy leads to more stable convergence and superior results. We use this by default.

## 5. Conclusion

In this work, we propose Dual Reciprocal Learning (DRL) framework for language-based human motion understanding and generation. By fostering the strong reciprocity relationship with collaborative dual learning, DRL reduces data dependency, enhances model performance, and shows great promise for real-world applications where labeled pairs are harder to acquire. We further propose a generalized protocol that extends existing evaluation to a general-domain cross-modal feature space, mitigating dataset-specific bias.

**Acknowledgement.** This work was supported by the National Science and Technology Major Project (No. 2023ZD0121300), Zhejiang Provincial Natural Science Foundation of China (No. LD25F020001), Fundamental Research Funds for the Central Universities (226-2025-00057), and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2025A02), China Postdoctoral Science Foundation (No. 2025T180421), and the Postdoctoral Fellowship Program of CPSF (No. GZC20251066).

## References

- [1] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics Auton. Syst.*, 2018. 1, 3
- [2] Jongmin Kim, Hoill Jung, MyungA Kang, and Kyungyong Chung. 3d human-gesture interface for fighting games using motion recognition sensor. *Wirel. Pers. Commun.*, 2016. 1
- [3] Gregory F Welch. History: The use of the kalman filter for human motion tracking in virtual reality. *Presence*, 2009. 1
- [4] Chen Liang, Wenguan Wang, and Yi Yang. Towards human-like virtual beings: Simulating human behavior in 3d scenes. In *ICCV*, 2025. 1
- [5] Yating Wei. [retracted] deep-learning-based motion capture technology in film and television animation production. *Sec. Commun. Netw.*, 2022. 1
- [6] Yusuke Goutsu and Tetsunari Inamura. Linguistic descriptions of human motion with generative adversarial seq2seq learning. In *ICRA*, 2021. 1, 3
- [7] 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. 1, 3, 5, 6, 7
- [8] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2024. 1, 3, 5, 6, 7
- [9] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *CVPR*, 2024. 1, 3, 5, 6, 7
- [10] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 1, 3
- [11] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, 2023. 3
- [12] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*. 3, 6
- [13] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024.
- [14] 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. *CVPR*, 2022. 1, 3, 6
- [15] 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*, 2023. 1, 3, 5, 6
- [16] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Att2m: Text-driven human motion generation with multi-perspective attention mechanism. In *ICCV*, 2023. 1, 3, 6
- [17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 1
- [18] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.
- [19] Wenguan Wang, Yi Yang, and Yunhe Pan. Visual knowledge in the big model era: Retrospect and prospect. *FITEE*, 2025. 1
- [20] 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*, 2022. 1, 2, 5, 6, 8
- [21] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 5, 7
- [22] 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. *NeurIPS*, 2023. 2, 5
- [23] 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. 2, 6
- [24] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*, 2019. 2
- [25] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *CVPR*, 2020.
- [26] Zhichao Zhai, Guikun Chen, Wenguan Wang, Dong Zheng, and Jun Xiao. Taga: Self-supervised learning for template-free animatable gaussian articulated model. In *CVPR*, 2025. 2
- [27] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, 2020. 3
- [28] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021.
- [29] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *ECCV*, 2022.
- [30] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In *ICME*, 2023.
- [31] Xiao Liu, Guangyi Chen, Yansong Tang, Guangrun Wang, Xiao-Ping Zhang, and Ser-Nam Lim. Language-free compositional action generation via decoupling refinement. In *IEEE ICASSP*, 2024.
- [32] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *AAAI*, 2020.
- [33] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2301.03949*, 2023.
- [34] Sumith Kulal, Jiayuan Mao, Alex Aiken, and Jiajun Wu. Programmatic concept learning for human motion description and synthesis. In *CVPR*, 2022. 3

- [35] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *ICLR*, 2020. 3
- [36] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020.
- [37] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.
- [38] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *TOMM*, 2022.
- [39] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, 2022.
- [40] Miki Okamura, Naruya Kondo, Tatsuki Fushimi Maki Sakamoto, and Yoichi Ochiai. Dance generation by sound symbolic words. *arXiv preprint arXiv:2306.03646*, 2023.
- [41] Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. In *ACM MM*, 2023.
- [42] Xin Gao, Li Hu, Peng Zhang, Bang Zhang, and Liefeng Bo. Dancemeld: Unraveling dance phrases with hierarchical latent codes for music-to-dance synthesis. *arXiv preprint arXiv:2401.10242*, 2023.
- [43] Qiaosong Qi, Le Zhuo, Aixi Zhang, Yue Liao, Fei Fang, Si Liu, and Shuicheng Yan. Diffdance: Cascaded human motion diffusion model for dance generation. In *ACM MM*, 2023.
- [44] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023.
- [45] Siqi Yang, Zejun Yang, and Zhisheng Wang. Longdancediff: Long-term dance generation with conditional diffusion model. *arXiv preprint arXiv:2308.11945*, 2023.
- [46] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *NeurIPS*, 2019.
- [47] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE TMM*, 2020.
- [48] Buyu Li, Yongchi Zhao, Zhelun Shi, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI*, 2022. 3
- [49] Qiuqing Lu, Yipeng Zhang, Mingjian Lu, and Vwani Roychowdhury. Action-conditioned on-demand motion generation. In *ACM MM*, 2022. 3
- [50] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, 2024. 3, 6
- [51] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: bidirectional autoregressive motion model. In *ECCV*, 2024. 3, 6
- [52] 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 *CVPR*, 2022. 3
- [53] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023. 3
- [54] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *NeurIPS*, 2016. 3
- [55] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. Dual supervised learning. In *ICML*, 2017. 3
- [56] Zhibing Zhao, Yingce Xia, Tao Qin, Lirong Xia, and Tie-Yan Liu. Dual learning: Theoretical study and an algorithmic extension. In *ACML*, 2020. 3
- [57] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *CVPR*, 2018. 3
- [58] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, 2019. 3
- [59] Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. Mutual learning for acoustic matching and dereverberation via visual scene-driven diffusion. In *ECCV*, 2025. 3
- [60] Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *CVPR*, 2022. 3
- [61] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 4, 5
- [62] 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*, 2021. 4, 5
- [63] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*. 5, 7
- [64] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 5
- [65] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5
- [66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 5
- [67] 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. 5
- [68] 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. *NeurIPS*, 2022. 5

- [69] 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. 5
- [70] 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. 5
- [71] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 5
- [72] 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. *NeurIPS*, 2017. 6
- [73] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. <https://github.com/Maluuba/nlg-eval>, 2017. 7
- [74] nlg-metricverse Contributors. nlg-metricverse. <https://github.com/disi-unibo-nlp/nlg-metricverse>, 2023. 7
- [75] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 7
- [76] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop Text Summarization Branches Out*, 2004. 7