

Tri-Modal Motion Retrieval by Learning a Joint Embedding Space

Kangning Yin¹ Shihao Zou² Yuxuan Ge¹ Zheng Tian^{1*}

¹ ShanghaiTech University ² Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

{yinkn2022, geyx2023, tianzheng}@shanghaitech.edu.cn, sh.zou@siat.ac.cn

Abstract

Information retrieval is an ever-evolving and crucial research domain. The substantial demand for high-quality human motion data especially in online acquirement has led to a surge in human motion research works. Prior works have mainly concentrated on dual-modality learning, such as text and motion tasks, but three-modality learning has been rarely explored. Intuitively, an extra introduced modality can enrich a model’s application scenario, and more importantly, an adequate choice of the extra modality can also act as an intermediary and enhance the alignment between the other two disparate modalities. In this work, we introduce LAVIMO (LAnguage-Video-MOTion alignment), a novel framework for three-modality learning integrating human-centric videos as an additional modality, thereby effectively bridging the gap between text and motion. Moreover, our approach leverages a specially designed attention mechanism to foster enhanced alignment and synergistic effects among text, video, and motion modalities. Empirically, our results on the HumanML3D and KIT-ML datasets show that LAVIMO achieves state-of-the-art performance in various motion-related cross-modal retrieval tasks, including text-to-motion, motion-to-text, video-to-motion and motion-to-video. Our project webpage can be found in <https://lavimo2023.github.io/LAVIMO/>.

1. Introduction

Recent technologies such as Generative Adversarial Networks (GANs) [9], diffusion models [16, 17, 41] and multi-modal models [34, 47] have achieved significant advancements. This progress has spurred a surge in human motion related research including motion generation [43, 48, 53] and motion retrieval [20, 28, 49]. Motion generation, while effective in generating realistic motion conditioned on textual data [14, 27, 43, 55], music [44, 59] and motion sequences [31, 38, 52, 56], it often struggles with generating diverse or contextually appropriate motions in complex

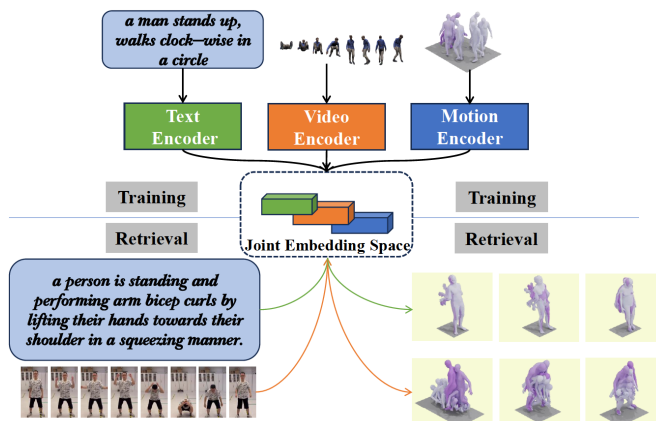


Figure 1. **Overview of LAVIMO.** During the training phase, the three modalities are processed through their distinct encoders. Subsequently, the resultant embeddings are aligned within a unified joint embedding space utilizing contrastive learning techniques. In the inference stage, the model is capable of accepting texts or videos as input queries, enabling the retrieval of corresponding motion data effectively.

scenarios and offers little controllability [58]. In contrast, motion retrieval completes motion acquirement through retrieving specific human motion sequences from a large dataset or database given queries commonly in forms of text, motion sequences, text, images, etc. It addresses the above challenges by leveraging external databases with high quality and diversity. Therefore, it is particularly valuable in industries with stringent requirements for precision, realism and controllability such as character animation [2, 18], virtual fitting rooms [15] and film production [26].

Many effective prior works in motion retrieval focus on dual-modality problems, especially text and motion related tasks given the generality of language. For example, TMR [28] utilizes contrastive learning to construct a joint embedding space for text-to-motion and motion-to-text retrieval. However, prior works often face a challenge that the spatial distance between two modalities such as text and motion being considerably vast [54], thus merging them into a unified embedding space often requires large datasets with high quality labels, which are also scarce. In comparison, the human-centric video modality serve as a com-

*Correspondence to Zheng Tian <tianzheng@shanghaitech.edu.cn>

pact, low-dimensional representation of intricate 3D motions [57], placing it closer in the spatial spectrum to motion. On the other hand, the availability of numerous video-text datasets bolsters the integration of the human-centric video and text modalities. Therefore, in this work, we take video as an intermediary modality to effectively narrow the spatial distance between each pair of the three modalities. Similar ideas can also be seen in other domain, for example, VALOR [3] leverages audio as an auxiliary modality to bolster the alignment between video and language. Building upon the integration of video into a cross-modal learning framework, we can also empower the model to execute video-to-motion retrieval task.

Conventional multi-modal learning frameworks often include a reconstruction branch to ensure that the information is not lost when translating from one modality to another. However, these frameworks tend to concentrate on data from only one specific modality, overlooking the distinct information inherent in other modalities. For instance, text contains specific details such as intentions and action paces, which are absent in video and motion modalities. Similarly, video offers unique visual information that is absent in other formats. In this work, we introduce an attention mechanism in which motion is utilized as queries. This mechanism is designed to extract relevant information from texts and videos, thereby addressing the limitations mentioned above. To the best of our knowledge, this paper first proposes to leverage video as an additional modality to enhance the alignment between the three modalities.

Our key contributions are summarized as follows: (i) We introduce LAnguage-Video-MOTion Alignment (LAV-IMO), a framework designed to cultivate a cohesive embedding space across the three aforementioned modalities. This architecture can accomplish both text-to-motion retrieval and video-to-motion retrieval tasks. (ii) We create a custom attention mechanism to merge the three modalities in our motion reconstruction branch, enabling us to incorporate valuable information from text and video when reconstructing motion sequences. This approach is intended to improve the alignment between the three modalities. (iii) In our efforts to enrich the available data, we augment both the HumanML3D and KIT datasets with RGB videos. This is achieved by animating and rendering avatars corresponding to specific motions. Such augmentation paves the way for more expansive and informed multi-modal research in human motion dynamics.

2. Related Work

Motion Generation. Motion generation is a popular topic in recent years. Existing frameworks for motion generation can be broadly categorized into two types. The first is a joint representation framework, in which text inputs are first processed by a pre-trained text encoder. Subsequently, these processed inputs are passed through a specially designed ar-

chitecture, such as a diffusion model, along with the original motion sequences. This process aims to effectively reconstruct motion sequences. The representative works in this category are MDM [43] and T2M-GPT [54]. The second so-called coordinated representation framework focuses on constructing a joint embedding space between texts and motions leveraging Auto-Encoders (AE) and Variational Auto-Encoders (VAE). The representative works in this category are MotionCLIP [42], TM2T [14] and TEMOS [27]. Our work is different from them as we incorporate a third modality and align the three modalities in a joint space.

Motion retrieval. Motion retrieval has gained popularity due to the rapid growth of motion capture data [6, 23, 50]. It offers an alternative method for obtaining motion sequences. Its primary goal is to identify and extract the motion sequence that exhibits the highest degree of similarity from an extensive motion database. Previous works such as DreCon [2] and PFNN [18] mainly focus on motion-to-motion retrieval, the specific motion is retrieved according to movement direction, heading direction, speed and locomotion style. However, motions are often represented parametrically, which can be challenging for individuals who are not experts in motion related research to comprehend. To address this limitation, TMR [28] initiates a novel branch termed as text-to-motion retrieval. This approach involves creating a cross-modal embedding space using contrastive learning. Additionally, TMR employs a negative filtering technique, ensuring that texts with similar meanings are considered as positive pairs. Nicola et al. [24] have also explored this field in their work. The primary goal of text-to-motion retrieval is to search for motion sequences in a database based on a specific text query. The performance of TMR has not reached its potential compared to that in the image and language domain. This is largely due to the lack of motion datasets. We incorporate an additional video modality to bridge the gap between text and motion, thereby reducing the reliance on extensive datasets.

Multi-Modal Learning. In the initial stages of multi-modal research, two common techniques are early fusion and late fusion [7, 10, 40]. In early fusion, features from different modalities are combined before being fed to a classifier. While in late fusion, features are processed separately, and their outputs are combined only in the final stages. In recent years, the field of multi-modal learning has experienced rapid advancements. Specifically, frameworks such as CLIP [32], DALL-E [33] and BERT [5] concentrate on elucidating the relationship between text and image. MV-GPT [39], Cap4Video [51] and EMCL [19] adeptly bridge the connection between text and video. Most of the works mentioned above utilize contrastive learning to build a joint embedding space between two modalities. Recent advancements in the video-language understanding domain start to leverage auxiliary modalities to enhance the

alignment between video and text. Among these, Yusuf et al. [1] effectively utilizes a teacher-student model and ranking loss to develop a cross-modal representation bridging text, image, and audio. Mithun et al. [25] skillfully integrates image appearance features, temporal dynamics, and audio cues from videos to enrich the information pool. VALOR [3], which introduces a vision-audio-language pre-training model stands out as a notable example.

3. Method

Our primary goal is to train a joint embedding space of three modalities, *i.e.*, text, video and motion, ideally allowing them to mutually represent each other. With this embedding space, we can accomplish various tasks involving text-to-motion retrieval or video-to-motion retrieval. In this section, we begin by providing a comprehensive definition of the three related modalities along with their distinct encoders in Sec. 3.1. Following that, we present the three-modality co-training process in Sec. 3.2 and the description of motion-related retrieval tasks in Sec. 3.3.

3.1. Modality Definitions and Model Architecture

Our extensive model architecture includes a motion encoder, a text encoder, and a video encoder. Each encoder independently extracts features from motion, text, and video, which are then aligned in a joint embedding space. To accelerate convergence, we initialized these encoders with pre-trained models.

Motion Encoder. The motion sequence comprises a series of 3D human poses $P = [P_1, P_2, \dots, P_{l_m}]$, where l_m represents the temporal length of a motion sequence and $P \in \mathbb{R}^{l_m \times c_m}$ with c_m being the feature dimension of a single pose. Each pose is parameterized using the SMPL model [22]. We follow the motion representation used in [13], which consists of joint velocities, local joint positions, 6D form local joint rotations and the foot contact label. We utilize the motion encoder from MotionCLIP [42]. In this process, the sequence P undergoes a linear projection to map into the latent space. A [CLS] token, representing the global feature of the motion sequence, is then appended at the beginning. This processed sequence is fed into the transformer encoder. We use the first position of the output as the representative motion feature, denoted as $e_m \in \mathbb{R}^{B \times C}$, where B is batch size and C is the latent dimension.

Text Encoder. The text description in our model is a sequence of words $T = [W_1, W_2, \dots, W_{l_t}]$ that succinctly summarize the contents of the motion, where l_t is the number of words. This description typically includes elements such as the target character (*e.g.*, “The man”), specific actions performed by the character (*e.g.*, “doing jumping jacks”), and stylistic attributes that characterize these actions (*e.g.*, “at a fast speed”, “counterclockwise”). Additionally, a single text description can involve multiple ac-

tions arranged in a sequence (*e.g.*, “The man walks forward at a rapid pace and then makes a right turn.”). We employ DistilBERT [37] as the foundational model for text processing. Textual tokens are generated through the DistilBERT tokenizer. Following the architecture of the motion encoder, a [CLS] token is appended at the beginning of the sequence. This entire sequence is then processed through the transformer encoder. The textual feature representation is extracted from the first output position, denoted as $e_t \in \mathbb{R}^{B \times C}$.

Video Encoder. A video is represented as a sequence of frames $I = [I_1, I_2, \dots, I_{l_v}]$, where we uniformly sample l_v frames from an RGB video as the input to avoid redundancy. The video encoder consists of a CLIP [34] image encoder and a temporal transformer. Initially, the video sequence is processed through the CLIP pre-trained image encoder to extract image features for each selected frame. Subsequently, considering the temporal aspects of the video, we input these image features into a temporal transformer. The video embedding is obtained by applying average pooling to the output of this temporal transformer, denoted as $e_v \in \mathbb{R}^{B \times C}$.

3.2. Cross-Modality Contrastive Learning

Multi-Modality Alignment. We employ contrastive learning to construct a fine-grained embedding space that bridges the three modalities. This alignment is achieved by minimizing the following loss function:

$$\mathcal{L}_{align} = \mathcal{L}_{align}^{mt} + \mathcal{L}_{align}^{mv} + \mathcal{L}_{align}^{tv}, \quad (1)$$

where \mathcal{L}_{align}^{xy} is the alignment loss between modality x and modality y .

Taking motion and text as an example, to maximize the proximity between positive samples while minimizing it between negative samples, we use the Kullback–Leibler (KL) divergence loss to establish our joint embedding space between text and motion:

$$\mathcal{L}_{align}^{mt} = KL(S_{pred}^{t2m}, S_{target}) + KL(S_{pred}^{m2t}, S_{target}^T), \quad (2)$$

where S_{pred}^{t2m} and S_{pred}^{m2t} represent the similarity matrix between text to motion and motion to text respectively. These can be calculated as follows:

$$S_{pred}^{t2m}(i, j) = \frac{\exp(\cos(e_t^i, e_m^j)/\tau)}{\sum_{j=1}^B \exp(\cos(e_t^i, e_m^j)/\tau)}, \quad (3)$$

$$S_{pred}^{m2t}(i, j) = \frac{\exp(\cos(e_m^i, e_t^j)/\tau)}{\sum_{j=1}^B \exp(\cos(e_m^i, e_t^j)/\tau)}, \quad (4)$$

where τ is the learnable temperature parameter, the indices (i, j) refer to the specific element within the respective matrix and the function $\cos(\cdot)$ represents the cosine similarity

between the two input embeddings:

$$\cos(e^i, e^j) = \frac{e^i \cdot e^{j\top}}{\|e^i\| \|e^j\|}. \quad (5)$$

Meanwhile, S_{target} is the target similarity matrix. However, considering that different text descriptions might convey similar meanings, defining S_{target} strictly as an identity matrix can pose issues. Such a definition would cause the model to incorrectly categorize texts with similar meanings as negative pairs, which could hinder the model’s efficacy in text-motion retrieval tasks. To address the aforementioned issue, we adopt the negative filtering technique proposed by TMR [28]. Specifically, we employ a pre-trained language model [21] to encode text labels into embeddings, which serve as the ground truth, denoted as \hat{e}_t . Subsequently, we calculate the cosine similarity between pairs of text embeddings within a single batch of size B . We define a threshold, ϵ , to represent the minimum similarity score required for a pair to be considered positive. Unlike the method used in TMR, instead of discarding scores above the threshold ϵ , we retain these scores in S_{target} . Therefore, we establish the target similarity matrix S_{target} with each element defined as:

$$S_{target}(i, j) = \begin{cases} \cos(\hat{e}_t^i, \hat{e}_t^j), & \text{if } \cos(\hat{e}_t^i, \hat{e}_t^j) \geq \epsilon, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Features Fusion. In this module, we incorporate a specially designed attention mechanism to blend features effectively during the motion reconstruction process. We treat motion embeddings as queries, which actively seek relevant information across textual and visual domains. The process can be expressed by:

$$\begin{aligned} \tilde{e}_m &= \text{Atten}(Q_{e_m}, K_{e_m}, V_{e_m}) + \text{Atten}(Q_{e_m}, K_{e_t}, V_{e_t}) \\ &+ \text{Atten}(Q_{e_m}, K_{e_f}, V_{e_f}), \end{aligned} \quad (7)$$

where $\text{Atten}(\cdot)$ is self-attention operations proposed in [45], and Q, K, V represent query, key and value, respectively. The subscripts e_m, e_t and e_v indicate the source of the input for these components. We feed the fused embedding \tilde{e}_m along with the duration of the motion sequence l_m into a transformer motion decoder to acquire the reconstructed motion sequence, denoted as P_{recon} . To evaluate the difference between the original and reconstructed motions, we utilize a Mean Squared Error (MSE) loss in training:

$$\mathcal{L}_{recon} = \text{MSE}(P, P_{recon}). \quad (8)$$

Our adoption of a customized attention mechanism is motivated by the need to bridge a gap in existing multi-modal alignment methods. These methods often incorporate a reconstruction branch to preserve the integrity of the learned

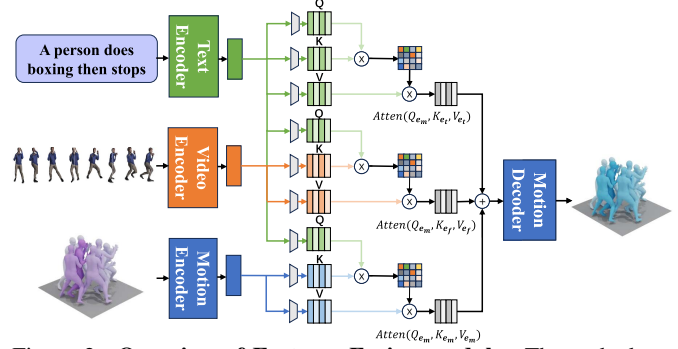


Figure 2. **Overview of Features Fusion module.** The embeddings for the text, video, and motion modalities are derived from their respective encoders. Subsequently, the motion embedding acts as a query to retrieve relevant information from the text and video, potentially compensating for any information that may be missing in the motion modality. The output of the attention mechanism is the weighted synthesis of the three modalities, which is then fed to the motion decoder for reconstruction.

representations in relation to the input data. The prior methods [34, 35] often focus on the information of a single modality, overlooking the rich contextual information available from other modalities. For instance, text frequently communicates subtle details that are not immediately apparent in motions or videos, such as intentions, explicit actions, or temporal nuances (e.g., “quickly” or “slowly”). In a similar pattern, videos offer visual context that might not be fully captured by text or motion. Thus, we draw inspiration from the query, key, and value framework commonly used in natural language processing [8, 46] to train our model for effective integration of crucial information from each modality. By extracting information from text and video modalities, our model can resolve ambiguities and fill in informational gaps within the motion modality. Our final training loss can be summarized as:

$$\mathcal{L} = \mathcal{L}_{align} + \lambda_{recon} \cdot \mathcal{L}_{recon}, \quad (9)$$

where λ_{recon} is the weight of the reconstruction loss.

3.3. Cross-Modalities Retrieval

Leveraging the learned joint embedding space across the three modalities, our method achieves various motion-related cross-modal retrieval tasks.

Retrieval Between Text and Motion. This task is proposed by TMR [28]. The objective is to search the most similar motion sequences in response to text queries, or conversely, to find text descriptions that best match given motion sequences. The text-to-motion retrieval task offers a unique advantage compared to text-to-motion generation, particularly in scenarios where high authenticity and efficiency are crucial, such as in animation and game production.

Retrieval Between Video and Motion. This task shares similarities with Human Pose Estimation (HPE) but also addresses several of its key limitations. While HPE has been a

Protocol	Methods	Text-to-motion retrieval						Motion-to-text retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
(a) All	TEMOS [27]	2.12	4.09	5.87	8.26	13.52	173.00	3.86	4.54	6.94	9.38	14.00	183.25
	Guo et al. [12]	1.80	3.42	4.79	7.12	12.47	81.00	2.92	3.74	6.00	8.36	12.95	81.50
	MotionCLIP [42]	2.33	5.85	8.93	12.77	18.14	103.00	5.12	6.97	8.35	12.46	19.02	91.42
	TMR [28]	5.68	10.59	14.04	20.34	30.94	28.00	9.95	12.44	17.95	23.56	32.69	28.50
	Ours(2-modal)	5.93	10.89	14.67	20.78	32.47	24.00	8.08	12.30	16.85	23.39	34.64	23.50
	Ours(3-modal)	6.37	11.84	15.60	21.95	33.67	24.00	9.72	13.33	18.73	25.00	36.55	22.50
(b) All with threshold	TEMOS [27]	5.21	8.22	11.14	15.09	22.12	79.00	5.48	6.19	9.00	12.01	17.10	129.0
	Guo et al. [12]	5.30	7.83	10.75	14.59	22.51	54.00	4.95	5.68	8.93	11.64	16.94	69.50
	MotionCLIP [42]	7.22	10.58	13.48	19.07	23.65	55.00	7.10	10.21	13.57	20.04	25.44	53.87
	TMR [28]	11.60	15.39	20.50	27.72	38.52	19.00	13.20	15.73	22.03	27.65	37.63	21.50
	Ours(2-modal)	11.32	15.01	20.91	27.93	40.97	17.00	12.95	17.56	22.69	29.05	39.58	19.00
	Ours(3-modal)	12.94	17.38	23.63	30.13	42.46	16.00	13.89	17.11	23.83	29.93	41.09	17.50
(c) Dissimilar subset	TEMOS [27]	20.00	33.00	37.00	47.00	62.00	6.00	24.00	30.00	39.00	47.00	62.00	6.74
	Guo et al. [12]	13.00	27.00	39.00	51.00	72.00	5.00	24.00	39.00	46.00	58.00	71.00	4.50
	MotionCLIP [42]	21.00	31.00	37.00	49.00	65.00	6.00	21.00	36.00	45.00	53.00	69.00	5.12
	TMR [28]	34.00	56.00	61.00	68.00	76.00	2.00	47.00	55.00	65.00	71.00	78.00	2.50
	Ours(2-modal)	51.00	63.00	69.00	76.00	80.00	1.50	52.00	65.00	70.00	77.00	80.00	1.00
	Ours(3-modal)	50.00	71.00	79.00	86.00	90.00	1.00	56.00	73.00	81.00	86.00	92.00	1.00
(d) Small batches [12]	TEMOS [27]	40.49	53.52	61.14	70.96	84.15	2.33	39.96	53.49	61.79	72.40	85.89	2.33
	Guo et al. [12]	52.48	71.05	80.65	89.86	96.58	1.39	52.00	71.21	81.11	89.87	96.78	1.38
	MotionCLIP [42]	46.24	60.25	68.93	80.47	91.35	1.88	44.76	56.81	65.22	77.83	90.19	2.03
	TMR [28]	67.16	81.32	86.81	91.43	95.36	1.04	67.97	81.20	86.35	91.70	95.27	1.03
	Ours(2-modal)	66.74	78.98	83.67	87.29	91.66	1.02	67.87	79.62	84.08	88.27	91.61	1.01
	Ours(3-modal)	68.58	81.04	85.02	88.77	92.58	1.01	68.64	81.06	85.52	88.76	92.82	1.01

Table 1. **Text-to-motion Retrieval on HumanML3D.** Presented here are our results on the text-to-motion retrieval tasks conducted on the HumanML3D dataset. The results indicate that our 3-modal version exceeds the performance of Guo et al. [12], TEMOS [27], MotionCLIP [42] and TMR [28]. Moreover, our 3-modal version outperforms our 2-modal version, showing that an extra modality can indeed enhance the alignment between texts and motions. The most notable results are emphasized in **bold**.

groundbreaking technique, it often lacks robustness, which can lead to issues such as jittering or floating feet. These problems can significantly compromise the realism of the motions predicted by HPE. In contrast, motions retrieved from large-scale motion capture datasets offer a more authentic and realistic experience. These motions are carefully captured from real-world movements, making them highly suitable for applications that demand high fidelity and an accurate representation of human motion.

4. Experiment

We carry out our experiments using the HumanML3D dataset [12] and the KIT-ML dataset [4]. The results demonstrate that our model outperforms previous methods in both the text-to-motion retrieval and video-to-motion retrieval tasks. Our method also exhibits generalization capabilities when applied to real-world, human-centric videos.

Datasets. HumanML3D [12] is the largest 3D human motion dataset with text descriptions, featuring 14,616 motions and 44,970 textual descriptions. The texts comprise 5,371 unique words. The motion data, originally from AMASS [23] and HumanAct12 [11], is downsampled to 20 FPS, cropped to 10 seconds if longer, and oriented to the Z+ direction. Each motion has a minimum of 3 descrip-

tions. The dataset is divided into 80% training, 5% validation, and 15% test sets. KIT-ML [29] is a human motion dataset with 3,911 sequences and 6,278 text annotations, containing a vocabulary of 1,623 unique words. Motions are downsampled to 12.5 FPS. Each sequence has 1 to 4 descriptions, averaging 8 words. The dataset also follows an 80% training, 5% validation, and 15% test split.

RGB Videos for KIT-ML and HumanML3D. Since the two datasets only contain motion capture data without any RGB videos, we propose to animate avatars and then render RGB videos to augment the video modality for both datasets. For each motion, an avatar is randomly picked from 13 different avatars, then animated and rendered to form its corresponding RGB videos of size 512×512 .

Evaluation Metrics Our evaluation of retrieval performance follows that of TMR [28], please turn to supplementary materials for more details.

4.1. Comparison to Prior Works

In this section, we compare our framework to the prior works on text-to-motion retrieval and video-to-motion retrieval tasks. For the tasks associated with text-to-motion retrieval, we introduce two distinct versions of our framework: the 2-modality version which is only trained on text

Protocol	Methods	Text-to-motion retrieval						Motion-to-text retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
(a) All	TEMOS [27]	7.11	13.25	17.59	24.10	35.66	24.00	11.69	15.30	20.12	26.63	36.39	26.50
	Guo et al. [12]	3.37	6.99	10.84	16.87	27.71	28.00	4.94	6.51	10.72	16.14	25.30	28.50
	MotionCLIP [42]	4.87	9.31	14.36	20.09	31.57	26.00	6.55	11.28	17.12	25.48	34.97	23.00
	TMR [28]	7.23	13.98	20.36	28.31	40.12	17.00	11.20	13.86	20.12	28.07	38.55	18.00
	Ours(2-modal)	8.59	15.04	21.09	32.23	46.09	13.00	11.72	17.19	23.63	32.81	48.83	12.00
	Ours(3-modal)	10.16	19.92	24.61	34.57	49.80	11.00	15.43	20.12	26.95	34.57	53.32	10.00
(b) All with threshold	TEMOS [27]	18.55	24.34	30.84	42.29	56.37	7.00	17.71	22.41	28.80	35.42	47.11	13.25
	Guo et al. [12]	13.25	22.65	29.76	39.04	49.52	11.00	10.48	13.98	20.48	27.95	38.55	17.25
	MotionCLIP [42]	13.79	23.08	31.45	42.93	53.01	9.00	13.24	22.11	29.53	38.06	50.23	10.00
	TMR [28]	24.58	30.24	41.93	50.48	60.36	5.00	19.64	23.73	32.53	41.20	53.01	9.50
	Ours(2-modal)	24.02	30.86	42.73	54.69	70.09	5.00	21.68	27.93	34.18	42.77	57.42	8.00
	Ours(3-modal)	30.86	41.80	48.63	59.96	74.22	4.00	25.98	31.25	38.28	45.70	63.09	6.50
(c) Dissimilar subset	TEMOS [27]	24.00	40.00	46.00	54.00	70.00	5.00	33.00	39.00	45.00	49.00	64.00	6.50
	Guo et al. [12]	16.00	29.00	36.00	48.00	66.00	6.00	24.00	29.00	36.00	46.00	66.00	7.00
	MotionCLIP [42]	19.00	33.00	41.00	50.00	69.00	6.00	28.00	36.00	43.00	48.00	65.00	7.00
	TMR [28]	26.00	46.00	60.00	70.00	83.00	3.00	34.00	45.00	60.00	69.00	82.00	3.50
	Ours(2-modal)	29.00	45.00	60.00	71.00	81.00	2.00	43.00	59.00	67.00	73.00	83.00	2.00
	Ours(3-modal)	30.00	49.00	63.00	73.00	84.00	3.00	48.00	60.00	66.00	76.00	82.00	2.00
(d) Small batches [12]	TEMOS [27]	43.88	58.25	67.00	74.00	84.75	2.06	41.88	55.88	65.62	75.25	85.75	2.25
	Guo et al. [12]	42.25	62.62	75.12	87.50	96.12	1.88	39.75	62.75	73.62	86.88	95.88	1.95
	MotionCLIP [42]	41.29	55.38	69.50	78.83	90.12	1.73	39.55	52.07	68.13	77.94	90.85	2.16
	TMR [28]	49.25	69.75	78.25	87.88	95.00	1.50	50.12	67.12	76.88	88.88	94.75	1.53
	Ours(2-modal)	53.96	76.42	82.05	89.66	95.22	1.38	58.58	75.05	81.84	89.68	93.86	1.23
	Ours(3-modal)	58.10	77.80	86.34	93.08	96.47	1.08	60.23	77.52	86.44	93.22	95.87	1.20

Table 2. **Text-to-motion Retrieval on KIT-ML.** We conduct further evaluations of both our 2-modal and 3-modal approaches using the KIT-ML dataset. The findings reveal that our 2-modal version significantly surpasses previous methodologies in performance. Moreover, our 3-modal version demonstrates an even greater extent of superiority over other existing methods. The most notable results are emphasized in **bold**.

and motion modalities, and the 3-modality version which is trained on all the 3 modalities. We compare our proposed methodologies against previous methods, including Guo et al. [12], TEMOS [27], MotionCLIP [42] and TMR [28]. The comparative results of these evaluations are tabulated in Table 1 and Table 2. Upon assessing performance on the HumanML3D dataset, it shows that our 3-modality framework surpasses the current state-of-the-art TMR across all evaluation protocols for both text-to-motion and motion-to-text retrieval tasks. Our three-modality framework also outperforms the two-modality one, demonstrating that the additional video modality can indeed enhance the alignment between text and motion modalities. On the KIT-ML dataset, we observe a even larger margin compared to HumanML3D dataset. In our new proposed video-to-motion retrieval tasks, we benchmark our work against MotionCLIP [42] and MotionSet [36]. MotionCLIP stands out as the sole work we have identified that leverages an extra modality to boost outcomes in the realm of human motion. However, MotionCLIP uses one randomly selected RGB frame as the input, rather than a sequence of video frames, ignoring temporal features of motion sequences in the video modality. MotionSet converts the original video

or MoCap clips to binary silhouette sets and extracting the motion features by a MotionSet network. Given that MotionCLIP is initially trained on the BABEL dataset [30] and MotionSet is initially trained on a self-made dataset, we re-train both the models on the HumanML3D and KIT-ML datasets for equitable comparison. Furthermore, considering MotionCLIP’s incorporation of images as an auxiliary modality, we conduct a fair comparison by sampling eight images and applying mean pooling to the model’s outputs. Our evaluations, as detailed in Table 3 and Table 4, reveal that our 3-modality framework significantly surpasses the performance of MotionCLIP and MotionSet on both the HumanML3D and KIT-ML datasets.

4.2. Qualitative Results

Figure 3 presents the visual results for the test subset within the HumanML3D dataset [12]. In the text-to-motion retrieval task, our framework is compared against the TMR method [28]. By selecting a textual description randomly from the database, our model demonstrates precision in retrieving the exact ground-truth motion in the rank1 position. Subsequent motions, ranking second and third, while not perfect matches, exhibit similar key motion elements,

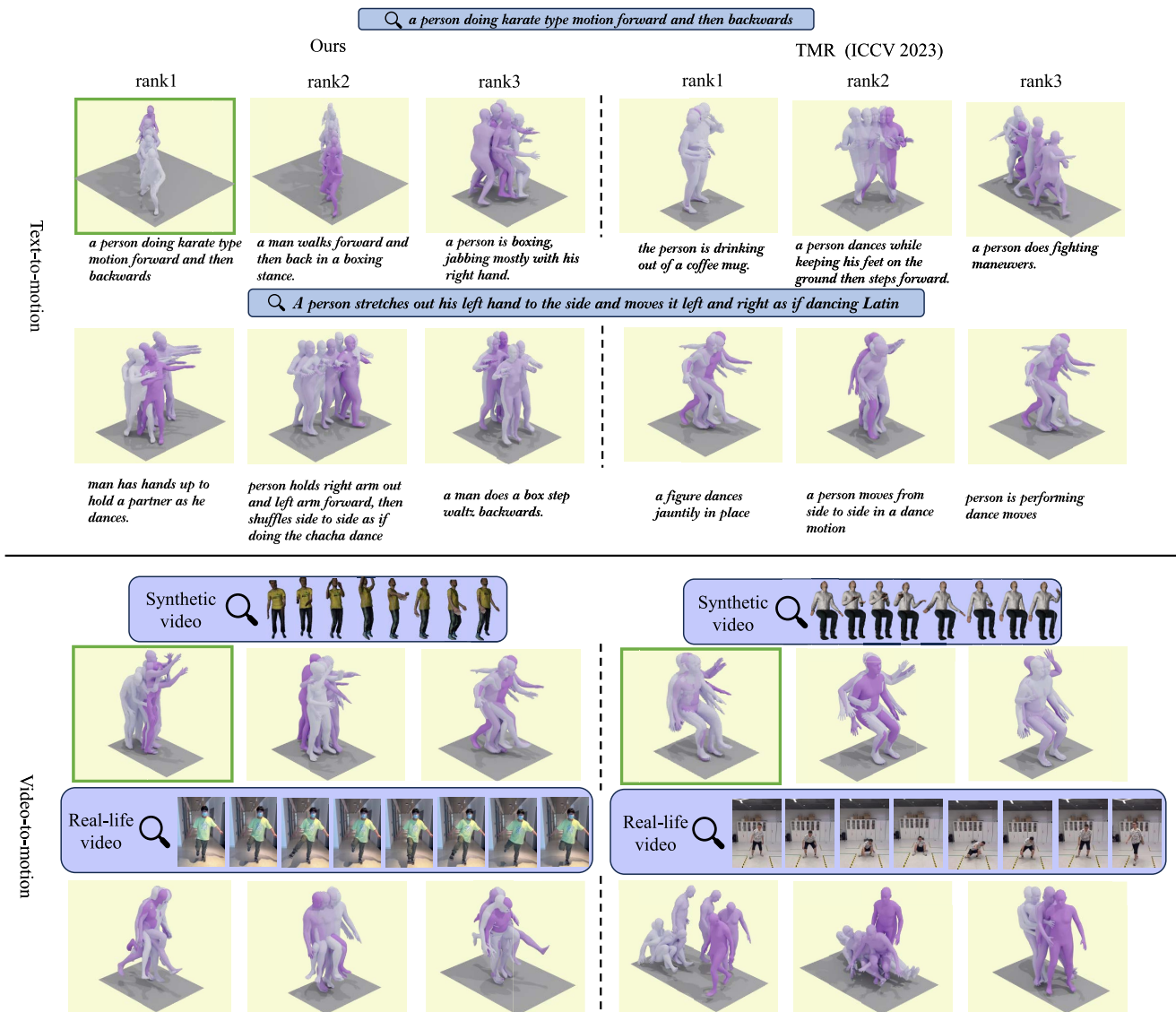


Figure 3. **Qualitative Comparison on the HumanML3D Dataset.** Our method successfully performs text-to-motion and video-to-motion retrieval tasks. For text-to-motion retrieval, we compare our results with TMR [28]. In the first row, using a random text from the test set, our method accurately retrieves the correct motion at rank 1, with similar motions such as “boxing” at ranks 2 and 3, resembling “karate type motion”. In contrast, TMR struggles, with only its rank 3 motion matching the ground truth. In the second row, when testing with a non-test set text involving “dance”, our model retrieves motions suggesting a “Latin dance”, more accurate than TMR’s less precise dance motions. For video-to-motion retrieval, in the third row, our model excels with test set videos, correctly retrieving ground-truth motions at rank 1 and similar motions at ranks 2 and 3. Furthermore, in the last row, when applied to real-life human-centric videos, our model shows strong generalization, retrieving motions closely matching the video content, such as “leg swinging” and “standing up and walking”.

including boxing and pacing activities. Moreover, when a text prompt describing a dance motion contained in the dataset is inputted, our model consistently identifies the top three corresponding motions, all indicative of Latin dance styles. In the video-to-motion retrieval task, we feed two rendered videos into the model. The first video features a character dribbling and shooting a basketball, while the second shows a character sitting and performing a “No” gesture. Our model retrieves the precise ground-truth motion in rank1 position, showing its effectiveness in the video-to-

motion retrieval task. In an additional assessment, when two real-life videos are fed into the model, the top three retrieved motions accurately reflect the actions depicted in the videos, demonstrating our model’s generalization ability to real-life videos.

4.3. User study: Video-to-Motion Retrieval

We create a small dataset comprising 100 real-life videos. The actions in these videos are performed in imitation of the motions found in the test sets of HumanML3D [12] and KIT-ML [29]. Specifically, we select 80 motions from Hu-

Protocol	Methods	Video-to-motion retrieval						Motion-to-video retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
(a) All	MotionCLIP [42]	4.96	9.40	12.48	17.46	26.46	43.00	3.52	7.28	9.74	13.82	22.34	58.00
	MotionSet [36]	8.26	18.69	35.83	48.91	56.32	16.00	10.17	21.33	37.39	51.25	58.97	13.00
	Ours(3-modal)	21.25	41.38	50.51	63.40	77.50	3.00	25.76	45.36	54.71	66.87	79.73	3.00
(b) All with threshold	MotionCLIP [42]	7.10	12.62	16.67	22.56	32.23	29.00	5.71	9.20	12.35	17.41	28.0	38.00
	MotionSet [36]	15.63	31.17	40.22	52.93	65.47	8.00	17.68	35.82	43.11	54.25	66.99	7.00
	Ours(3-modal)	30.42	46.97	56.82	68.77	81.41	3.00	35.43	52.73	61.93	72.95	84.54	2.00
(C)Dissimilar subset	MotionCLIP [42]	42.00	61.00	72.00	80.00	88.00	2.00	43.00	65.00	69.00	76.00	86.00	2.00
	MotionSet [36]	58.00	69.00	81.00	89.00	97.00	2.00	59.00	73.00	86.00	90.00	98.00	2.00
	Ours(3-modal)	74.00	93.00	98.00	100.00	100.00	1.00	79.00	99.00	100.00	100.00	100.00	1.00

Table 3. **Video-to-motion Retrieval on HumanML3D.** We assess the video-to-motion retrieval task using the HumanML3D datasets. Our approach surpasses MotionCLIP [42] and MotionSet [36] across all the evaluation protocols in the table. The most notable results are emphasized in **bold**.

Protocol	Methods	Video-to-motion retrieval						Motion-to-video retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
(a) All	MotionCLIP [42]	3.71	6.45	8.79	11.52	17.97	52.50	4.49	7.23	9.96	12.11	21.68	51.00
	MotionSet [36]	9.15	23.96	35.27	46.18	61.31	13.00	11.08	24.27	35.03	45.29	60.53	13.00
	Ours(3-modal)	18.75	35.16	42.58	55.86	73.24	4.00	23.05	35.55	44.14	56.25	72.07	4.00
(b) All with threshold	MotionCLIP [42]	8.98	12.11	14.84	19.53	27.73	27.00	9.18	12.50	17.19	22.27	33.98	21.00
	MotionSet [36]	20.17	35.21	43.38	55.39	69.42	6.00	24.57	37.66	48.29	59.72	70.71	5.00
	Ours(3-modal)	36.91	49.80	60.94	70.70	84.57	3.00	41.41	52.73	63.28	75.98	86.13	2.00
(C)Dissimilar subset	MotionCLIP [42]	15.00	23.00	31.00	40.00	59.00	8.00	13.00	21.00	25.00	41.00	63.00	6.00
	MotionSet [36]	40.00	58.00	73.00	85.00	89.00	2.00	40.00	60.00	73.00	82.00	87.00	2.00
	Ours(3-modal)	66.00	84.00	90.00	96.00	99.00	1.00	66.00	87.00	93.00	98.00	100.00	1.00

Table 4. **Video-to-motion Retrieval on KIT-ML.** We assess the video-to-motion retrieval task using the KIT-ML datasets. Our findings align with those obtained from the HumanML3D, indicating that our framework significantly outperforms the performance of MotionCLIP [42] and MotionSet [36] by a considerable margin.

manML3D and 20 from KIT-ML for imitation. To enhance diversity, we engage 5 performers, each with a distinctly different body type. We instruct users to view a real-life video and then choose the most similar video from four provided videos. Among these four videos, one is the top-ranked motion retrieved by our framework, while the other three are randomly selected from the corresponding dataset. Additionally, we offer a “None of the above” option for users in case they believe the real-life video does not closely match any of the provided options. The results show that our framework successfully retrieves 3D motions that are subjectively similar for 68.5% of the videos, demonstrating its effective generalization capabilities to real-life video scenarios. Our user study webpage can be found in <https://lavimo2023.github.io/User-Study-LAVIMO/>

4.4. Limitations

Although LAVIMO demonstrates enhanced efficacy in both text-to-motion and video-to-motion retrieval tasks, there are limitations that future research should consider addressing. Firstly, the video modality in our approach is derived from animating and rendering avatars to match specific motions, which inherently deviates from authentic human-centric videos. While our framework exhibits a degree of generalization to real-life video, the substitution of ren-

dered footage with actual human-centric videos may yield improvements. Secondly, despite outperforming existing benchmarks, our model does not yet match the precision achieved in video and image retrieval tasks. This discrepancy primarily stems from the limited availability of comprehensive motion datasets, which constrains the broader application potential of the motions retrieved, including their integration into the motion generation pipeline.

5. Conclusion and Discussion

In this paper, we introduce LAVIMO, a Unified Language-Video-Motion Alignment framework. By employing contrastive learning, we effectively integrate three distinct modalities into a cohesive embedding space. Our approach surpasses previous methodologies in text-to-motion retrieval tasks, as evidenced by our results on the HumanML3D and KIT-ML datasets. Furthermore, we propose a novel video-to-motion retrieval task, which may serve as an adjunct to existing video pose estimation techniques. Additionally, we demonstrate the capacity of our framework to generalize effectively to real-life human-centric video.

Acknowledgement

This work is supported by Shanghai Sailing Program (23YF1427600).

References

- [1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017. 3
- [2] Kevin Bergamin, Simon Clavet, Daniel Holden, and James Richard Forbes. Drecon: data-driven responsive control of physics-based characters. *ACM TOG*, 2019. 1, 2
- [3] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 2, 3
- [4] Simon Clavet. Motion matching and the road to next-gen animation. In *Proc. of GDC*, 2016. 5
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018. 2
- [6] Adso Fern'andez-Baena, Antonio Susin, and Xavier Lligadas. Biomechanical validation of upper-body and lower-body joint movements of kinect motion capture data for rehabilitation treatments. In *2012 fourth international conference on intelligent networking and collaborative systems*, 2012. 2
- [7] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetsche. Early vs late fusion in multimodal convolutional neural networks. In *FUSION*, 2020. 2
- [8] Hongyu Gao, Chao Zhu, Mengyin Liu, Weibo Gu, Hongfa Wang, Wei Liu, and Xu-Cheng Yin. Calic: Accurate and efficient image-text retrieval via contrastive alignment and visual contexts modeling. In *ACM MM*, 2022. 4
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 1
- [10] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics*, 2005. 2
- [11] 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. 5
- [12] 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. 5, 6, 7
- [13] 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. 3
- [14] 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, 2
- [15] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1
- [16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1
- [18] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM TOG*, 2017. 1, 2
- [19] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *NeurIPS*, 2022. 2
- [20] Mubbasir Kapadia, I-kao Chiang, Tiju Thomas, Norman I Badler, and Joseph T Kider Jr. Efficient motion retrieval in large motion databases. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2013. 1
- [21] Patrick S. H. Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. PAQ: 65 million probably-asked questions and what you can do with them. *CoRR*, 2021. 4
- [22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 3
- [23] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 5
- [24] Nicola Messina, Jan Sedmidubsky, Fabrizio Falchi, and Tomás Rebok. Text-to-motion retrieval: Towards joint understanding of human motion data and natural language. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2420–2425, 2023. 2
- [25] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 3
- [26] Markus Mühling, Nikolaus Korfhage, Eric Müller, Christian Otto, Matthias Springstein, Thomas Langelage, Uli Veith, Ralph Ewerth, and Bernd Freisleben. Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications*, 2017. 1
- [27] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 1, 2, 5, 6
- [28] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 1, 2, 4, 5, 6, 7
- [29] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 2016. 5, 7
- [30] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, 2021. 6
- [31] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *CVPR*, 2023. 1
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askill, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, 2021. [2](#)
- [33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, 2021. [2](#)
- [34] 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. [1](#), [3](#), [4](#)
- [35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. [4](#)
- [36] Tingxin Ren, Wei Li, Zifei Jiang, Xueqing Li, Yan Huang, and Jingliang Peng. Video-based human motion capture data retrieval via motionset network. *IEEE Access*, 8:186212–186221, 2020. [6](#), [8](#)
- [37] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, 2019. [3](#)
- [38] Ayumu Sasagawa, Sho Sakaino, and Toshiaki Tsuji. Motion generation using bilateral control-based imitation learning with autoregressive learning. *IEEE Access*, 2021. [1](#)
- [39] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 2022. [2](#)
- [40] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *ACM MM*, 2005. [2](#)
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [1](#)
- [42] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 2022. [2](#), [3](#), [5](#), [6](#), [8](#)
- [43] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. [1](#), [2](#)
- [44] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023. [1](#)
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. [4](#)
- [46] Quan Wang, Pingping Huang, Haifeng Wang, Songtai Dai, Wenbin Jiang, Jing Liu, Yajuan Lyu, Yong Zhu, and Hua Wu. Coke: Contextualized knowledge graph embedding. *arXiv preprint arXiv:1911.02168*, 2019. [4](#)
- [47] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey, 2023. [1](#)
- [48] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *ICCV*, 2023. [1](#)
- [49] Zhao Wang, Yinfu Feng, Tian Qi, Xiaosong Yang, and Jian J Zhang. Adaptive multi-view feature selection for human motion retrieval. *Signal Processing*, 2016. [1](#)
- [50] Xiaolin Wei, Peizhao Zhang, and Jinxiang Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM TOG*, 2012. [2](#)
- [51] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, pages 10704–10713, 2023. [2](#)
- [52] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*, 2019. [1](#)
- [53] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. [1](#)
- [54] 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. *arXiv preprint arXiv:2301.06052*, 2023. [1](#), [2](#)
- [55] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. [1](#)
- [56] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *CVPR*, 2020. [1](#)
- [57] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 2023. [2](#)
- [58] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiabin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey, 2023. [1](#)
- [59] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM TOMM*, 2022. [1](#)