MotionCLIP: Exposing Human Motion Generation to CLIP Space

Guy Tevet*, Brian Gordon*, Amir Hertz,
Amit H. Bermano, and Daniel Cohen-Or

Tel Aviv University, Israel
{guytevet, briangordon}@mail.tau.ac.il

Abstract. We introduce MotionCLIP, a 3D human motion auto-encoder featuring a latent embedding that is disentangled, well behaved, and supports highly semantic textual descriptions. MotionCLIP gains its unique power by aligning its latent space with that of the Contrastive Language-Image Pre-training (CLIP) model. Aligning the human motion manifold to CLIP space implicitly infuses the extremely rich semantic knowledge of CLIP into the manifold. In particular, it helps continuity by placing semantically similar motions close to one another, and disentanglement, which is inherited from the CLIP-space structure. MotionCLIP comprises a transformer-based motion auto-encoder, trained to reconstruct motion while being aligned to its text label's position in CLIP-space. We further leverage CLIP's unique visual understanding and inject an even stronger signal through aligning motion to rendered frames in a self-supervised manner. We show that although CLIP has never seen the motion domain, MotionCLIP offers unprecedented text-to-motion abilities, allowing out-of-domain actions, disentangled editing, and abstract language specification. For example, the text prompt "couch" is decoded into a sitting down motion, due to lingual similarity, and the prompt "Spiderman" results in a web-swinging-like solution that is far from seen during training. In addition, we show how the introduced latent space can be leveraged for motion interpolation, editing and recognition. [1]

1 Introduction

Human motion generation includes the intuitive description, editing, and generation of 3D sequences of human poses. It is relevant to many applications that require virtual or robotic characters. Motion generation is, however, a challenging task. Perhaps the most challenging aspect is the limited availability of data, which is expensive to acquire and to label. Recent years have brought larger sets of motion capture acquisitions [29], sometimes sorted by classes [25, 21] or even labeled with free text [37, 35]. Yet, it seems that while this data may span a significant part of human motion, it is not enough for machine learning algorithms

[1] See our project page: https://guytevet.github.io/motionclip-page/

* The authors contributed equally

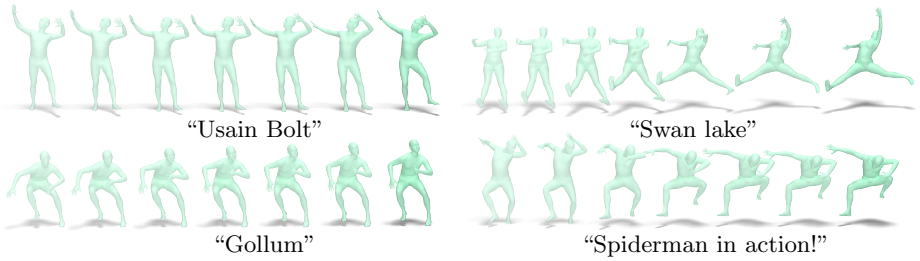

Fig. 1. Motions generated by MotionCLIP conditioned on different cultural references. MotionCLIP exploits the rich knowledge encapsulated in pre-trained language-images model (CLIP) and projects the human motion manifold over its latent space.

to understand the semantics of the motion manifold, and it is definitely not descriptive enough for natural language usage. Hence, neural models trained using labeled motion data [2, 24, 48, 33, 28] do not generalize well to the full richness of the human motion manifold, nor to the natural language describing it.

In this work, we introduce MotionCLIP, a 3D motion auto-encoder that induces a latent embedding that is disentangled, well behaved, and supports highly semantic and elaborate descriptions. To this end, we employ CLIP [38], a large scale visual-textual embedding model. Our key insight is that even though CLIP has not been trained on the motion domain what-so-ever, we can inherit much of its latent space's virtue by enforcing its powerful and semantic structure onto the motion domain. To do this, we train a transformer-based [43] auto-encoder that is aligned to the latent space of CLIP, using existing motion textual labels. In other words, we train an encoder to find the proper embedding of an input sequence in CLIP space, and a decoder that generates the most fitting motion to a given CLIP space latent code. To further improve the alignment with CLIP-space, we also leverage CLIP's visual encoder, and synthetically render frames to guide the alignment in a self-supervised manner (see Figure 2). As we demonstrate, this step is crucial for out-of-domain generalization, since it allows finer-grained description of the motion, unattainable using text.

The merit of aligning the human motion manifold to CLIP space is two-fold: First, combining the geometric motion domain with lingual semantics benefits the semantic description of motion. As we show, this benefits tasks such as text-to-motion and motion style transfer. More importantly however, we show that this alignment benefits the motion latent space itself, infusing it with semantic knowledge and inherited disentanglement. Indeed, our latent space demonstrates unprecedented compositionality of independent actions, semantic interpolation between actions, and even natural and linear latent-space based editing.

As mentioned above, the textual and visual CLIP encoders offer the semantic description of motion. In this aspect, our model demonstrates never-before-seen capabilities for the field of motion generation. For example, motion can be specified using arbitrary natural language, through abstract scene or intent descriptions instead of the motion directly, or even through pop-culture references. For

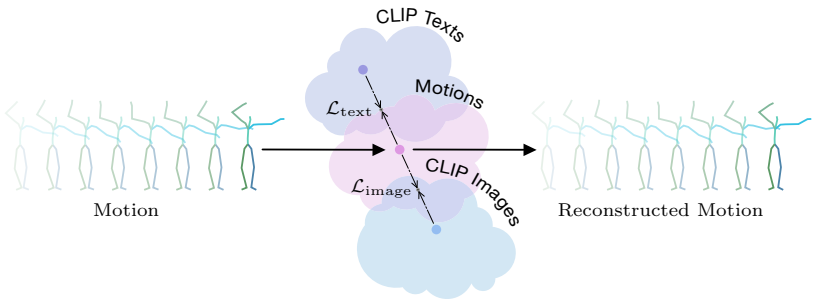

Fig. 2. MotionCLIP overview. A motion auto-encoder is trained to simultaneously reconstruct motion sequences while aligning their latent representation with corresponding texts and images representations in CLIP space.

example, the CLIP embedding for the phrase "wings" is decoded into a flapping motion like a bird, and "Williams sisters" into a tennis serve, since these terms are encoded close to motion seen during training, thanks to CLIP's semantic understanding. Through the compositionality induced by the latent space, the aforementioned process also yields clearly unseen motions, such as the iconic web-swinging gesture that is produced for the input "Spiderman" (see this and other culture references in Figure 1). Our model also naturally extents to other downstream tasks. In this aspect, we depict motion interpolation to depict latent smoothness, editing to demonstrate disentanglement, and action recognition to point out the semantic structure of our latent space. For all these applications, we show comparable or preferable results either through metrics or a user study, even though each task is compared against a method that was designed especially for it. Using the action recognition benchmark, we also justify our design choices with an ablation study.

2 Related Work

2.1 Guided Human Motion Generation

One means to guide motion generation is to condition on another domain. An immediate, but limited, choice is conditioning on *action* classes. ACTOR [33] and Action2Motion [14] suggested learning this multi-modal distribution from existing action recognition datasets using Conditional Variational-Autoencoder(CVAE) [42] architectures. MUGL [28] model followed with elaborated Conditional Gaussian-Mixture-VAE [6] that supports up to 120 classes and multi-person generation, based on the NTU-RGBD-120 dataset [25].

Motion can be conditioned on other domains. For example, recent works [23, 3] generated dance moves conditioned on music and the motion prefix. Edwards et al. [8] generated facial expressions to fit a speaking audio sequence.

A more straightforward approach to control motion is using another motion. In particular, for style transfer applications. Holden et al. [18] suggested

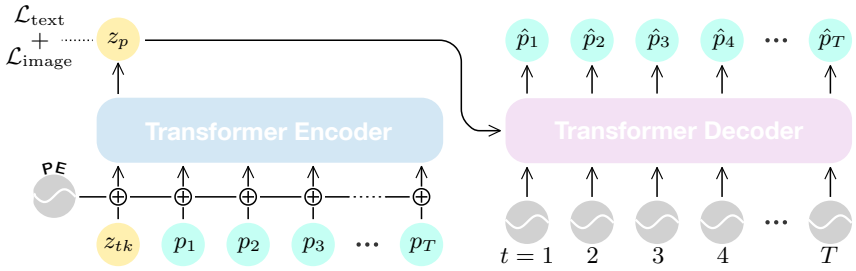

Fig. 3. Motion Auto-Encoder. A transformer encoder is trained to project a motion sequence $p_{1:T}$ into a latent vector z_p in CLIP latent space. Simultaneously, a transformer decoder is trained to recover the motion by attending to z_p.

to code style using the latent code's Gram matrix, inspired by Gatys et al. [12]. Aberman et al. [1] injected style attributes using a dedicated temporal-invariant AdaIN layer [20]. Recently, Wen et al. [47] encoded style in the latent code of Normalizing Flow generative model [7]. We show that MotionCLIP also encodes style in its latent representation, without making any preliminary assumptions or using a dedicated architecture.

2.2 Text-to-Motion

The KIT dataset[35] provides about 11 hours of motion capture sequences, each sequence paired with a sentence explicitly describing the action performed. KIT sentences describe the action type, direction and sometimes speed, but lacks details about the style of the motion, and not including abstract descriptions of motion. Current text-to-motion research is heavily based on KIT. Plappert et al. [36] learned text-to-motion and motion-to-text using seq2seq RNN-based architecture. Yamada et al. [48] learned those two mappings by simultaneously training text and motion auto-encoders while binding their latent spaces using text and motion pairs. Lin et al. [24] further improved trajectory prediction by adding a dedicated layer. Ahuja et al. [2] introduced JL2P model, which got improved results with respect to nuanced concepts of the text, namely velocity, trajectory and action type. They learned joint motion-text latent space and apply training curriculum to ease optimization. Concurrent to our work, Petrovich et al. [34] and Guo et al. [13] encourage diverse generation using VAE based models [22], yet not generalizing outside of the limited available data.

More recently, BABEL dataset [37] provided per-frame textual labels ordered in 260 classes to the larger AMASS dataset [29], including about 40 hours of motion capture. Although providing explicit description of the action, often lacking any details besides the action type, this data spans a larger variety of human motion. MotionCLIP overcomes the data limitations by leveraging out-of-domain knowledge using CLIP [38].

2.3 CLIP aided Methods

Neural networks have successfully learned powerful latent representations coupling natural images with natural language describing it [17, 39]. A recent example is CLIP[38], a model coupling images and text in deep latent space using a constructive objective[16, 4]. By training over hundred millions of images and their captions, CLIP gained a rich semantic latent representation for visual content. This expressive representation enables high quality image generation and editing, controlled by natural language [31, 11, 10]. Even more so, this model has shown that connecting the visual and textual worlds also benefits purely visual tasks [44], simply by providing a well-behaved, semantically structured, latent space.

Closer to our method are works that utilize the richness of CLIP outside the imagery domain. In the 3D domain, CLIP's latent space provides a useful objective that enables semantic manipulation [40, 30, 45] where the domain gap is closed by a neural rendering. CLIP is even adopted in temporal domains [15, 27, 9] that utilize large datasets of video sequences that are paired with text and audio. Unlike these works that focus on classification and retrieval, we introduce a generative approach that utilizes limited amount of human motion sequences that are paired with text.

More recently, CLIP was used for motion applications. CLIP-Actor [49] is using CLIP representation to query motions from BABEL. The motions in that case are represented by their attached textual labels. AvatarCLIP [19] used CLIP as a loss term for direct motion optimization, where the motion is represented by rendered poses. Contrary to their claim, MotionCLIP shows that motions can be successfully encoded-to and decoded-from CLIP space.

3 Method

Our goal is learning a semantic and disentangled motion representation that will serve as a basis for generation and editing tasks. To this end, we need to learn not only the mapping to this representation (encoding), but also the mapping back to explicit motion (decoding).

Our training process is illustrated in Figure 2. We train a transformer-based motion auto-encoder, while aligning the latent motion manifold to CLIP joint representation. We do so using (i) a *Text Loss*, connecting motion representations to the CLIP embedding of their text labels, and (ii) an *Image Loss*, connecting motion representations to CLIP embedding of rendered images that depict the motion visually.

At inference time, semantic editing applications can be performed in latent space. For example, to perform style transfer, we find a latent vector representing the style, and simply add it to the content motion representation and decode the result back into motion. Similarly, to classify an action, we can simply encode it into the latent space, and see to which of the class text embedding it is closest. Furthermore, we use the CLIP text encoder to perform text-to-motion

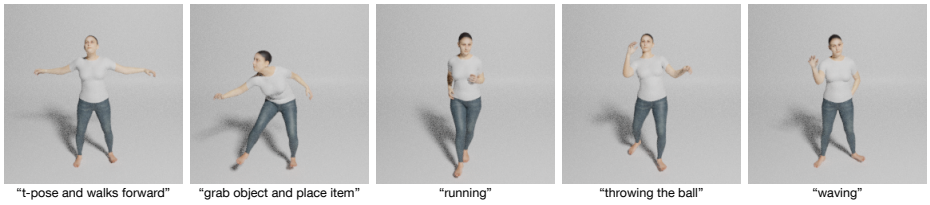

Fig. 4. A sample of the rendered frames and their text description used during training.

- An input text is decoded using the text encoder then directly decoded by our motion decoder. The implementation of these and other applications is detailed in Section 4.

We represent motion sequences using the SMPL body model [26]. A sequence of length T denoted $p_{1:T}$ such that $p_i \in \mathbb{R}^{24\times 6}$ defines orientations in 6D representation[50] for global body orientation and 23 SMPL joints, at the i^{th} frame. The mesh vertices locations $v_{1:T}$ are calculated according to SMPL specifications with $\beta = 0$ and a neutral-gender body model following Petrovich et al. [33].

To project the motion manifold into the latent space, we learn a transformer-based auto-encoder [43], adapted to the motion domain [33, 46, 23]. MotionCLIP's architecture is detailed in Figure 3.

Transformer Encoder. E, Maps a motion sequence $p_{1:T}$ to its latent representation z_p. The sequence is embedded into the encoder's dimension by applying linear projection for each frame separately, then adding standard positional embedding. The embedded sequence is the input to the transformer encoder, together with additional learned prefix token z_{tk}. The latent representation, z_p is the first output (the rest of the sequence is dropped out). Explicitly, $z_p = E(z_{tk}, p_{1:T})$.

Transformer Decoder. D, predicts a motion sequence $\hat{p}_{1:T}$ given a latent representation z_p. This representation is fed to the transformer as key and value, while the query sequence is simply the positional encoding of $1:T$. The transformer outputs a representation for each frame, which is then mapped to pose space using a linear projection. Explicitly, $\hat{p}_{1:T} = D(z_p)$. We further use a differentiable SMPL layer to get the mesh vertices locations, $\hat{v}_{1:T}$.

Losses. This auto-encoder is trained to represent motion via reconstruction $L2$ losses on joint orientations, joint velocities and vertices locations. Explicitly,

$$\begin{aligned}\mathcal{L}_{\text{recon}} = &\frac{1}{|p|T}\sum_{i=1}^{T}\|p_i - \hat{p}_i\|^2 + \frac{1}{|v|T}\sum_{i=1}^{T}\|v_i - \hat{v}_i\|^2 \\ &+\frac{1}{|p|(T-1)}\sum_{i=1}^{T-1}\|(p_{i+1} - p_i) - (\hat{p}_{i+1} - \hat{p}_i)\|^2\end{aligned} \tag{1}$$

Given text-motion and image-motion pairs, $(p_{1:T}, t)$, $(p_{1:T}, s)$ correspondingly, we attach the motion representation to the text and image representations using cosine distance,

$$\mathcal{L}_{\text{text}} = 1 - \cos(CLIP_{\text{text}}(t), z_p) \tag{2}$$

and

$$\mathcal{L}_{\text{image}} = 1 - \cos(CLIP_{\text{image}}(s), z_p) \tag{3}$$

The motion-text pairs can be derived from labeled motion dataset, whereas the images can be achieved by rendering a single pose from a motion sequence, to a synthetic image s, in an unsupervised manner (More details in Section 4).

Overall, the loss objective of MotionCLIP is defined,

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{image}}\mathcal{L}_{\text{image}} \tag{4}$$

4 Results

To evaluate MotionCLIP, we consider its two main advantages. In Section 4.2, we inspect MotionCLIP's ability to convert text into motion. Since the motion's latent space is aligned to that of CLIP, we use CLIP's pretrained text encoder to process input text, and convert the resulting latent embedding into motion using MotionCLIP's decoder. We compare our results to the state-of-the-art and report clear preference for both seen and unseen generation. We also show comparable performance to state-of-the-art style transfer work simply by adding the style as a word to the text prompt. Lastly, we exploit CLIP expert lingual understanding to convert abstract text into corresponding, and sometimes unexpected, motion.

In Sections 4.3 and 4.4 we focus on the resulting auto-encoder, and the properties of its latent-space. We inspect its smoothness and disentanglement by (1) conducting ablation study using established quantitative evaluation, and (2) demonstrating various applications. Smoothness is shown through well-behaved interpolations, even between distant motion. Disentanglement is demonstrated using latent space arithmetic; by adding and subtracting various motion embeddings, we achieve compositionality and semantic editing. Lastly, we leverage our latent structure to perform action recognition over the trained encoder. The latter setting is also used for ablation study. In the following, we first lay out the data used, and other general settings.

4.1 General Settings

We train our model on the BABEL dataset [37]. It comprises about 40 hours of motion capture data, represented with the SMPL body model [26]. The motions are annotated with per-frame textual labels, and is categorized into one of 260 action classes. We down sample the data to 30 frames per-second and cut it into sequences of length 60. We get a single textual label per sequence by listing all actions in a given sequence, then concatenating them to a single string. Finally,

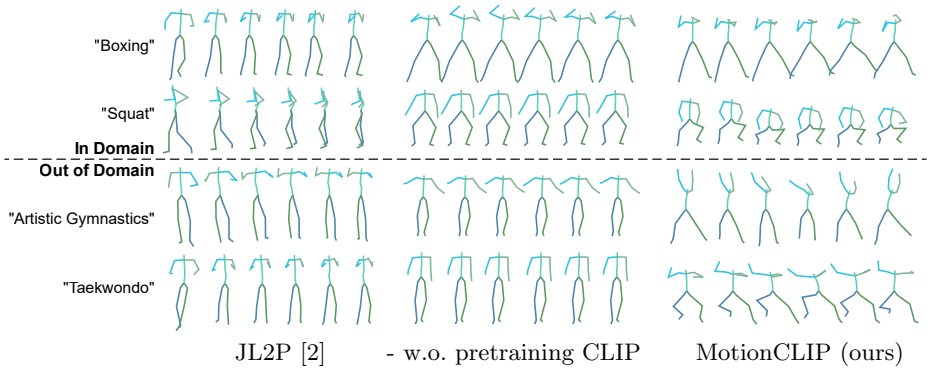

Fig. 5. In- and Out-of-domain Qualitative results for MotionCLIP with and without CLIP pretraining. MotionCLIP (right) performs better for in-domain motions than out-of-domain, and in any case better than JL2P and MotionCLIP ablated variant.

we choose for each motion sequence a random frame to be rendered using the *Blender* software and the SMPL-X add-on [32] (See Figure 4). This process outputs triplets of (motion, text, synthetic image) which are used for training.

We train a transformer auto-encoder with 8 layers for each encoder and decoder as described in Section 3. We align it with the *CLIP-ViT-B/32* frozen model. Out of the data triplets, the text-motion pairs are used for the *text loss* and image-motion pairs for the *image loss*. Both λ values are set to 0.01 throughout our experiments. [2]

4.2 Text-to-Motion

Text-to-motion is performed at inference time, using the CLIP text encoder and MotionCLIP decoder, without any further training. Even though not directly trained for this task, MotionCLIP shows unprecedented performance in text-to-motion, dealing with explicit descriptions, subtle nuances and abstract language.

Actions. We start by demonstrating the capabilities of MotionCLIP to generate explicit actions - both seen and unseen in training. We compare our model to JL2P [2] trained on BABEL and two ablated variants of MotionCLIP:(1) without CLIP pre-training (in this case, the text encoder is trained from scratch together with MotionCLIP, as in JL2P) and (2) without CLIP image loss (i.e. using text loss only). We distinguish between in-domain and out-of-domain actions by conducting a user study[3] using two different text sets: (1) The *in-domain set* comprises of BABEL-60 class names. (2) the *Out-of-domain set* includes textual labels that do not appear in any of the training labels. We construct this set from the list of Olympic sports that are disjoint to BABEL. Figure 6 shows that MotionCLIP is clearly preferred by the users over JL2P and the

[2] https://github.com/GuyTevet/MotionCLIP
[3] 30 unique users, each was asked 12 questions.

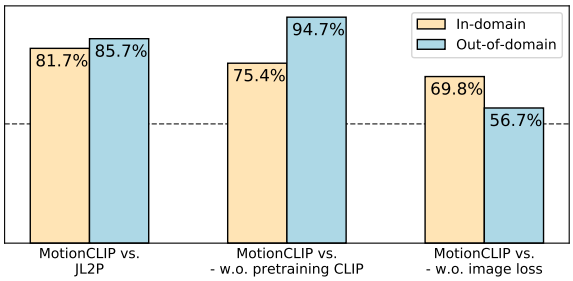

Fig. 6. Action generation from text - user study. The bars depict MotionCLIP's preference score vs. each of the other models (when compared side-by-side). The dashed line marks 50% (i.e. equally preferred models). MotionCLIP is clearly preferred by the users over JL2P [2] and our two ablated variants.

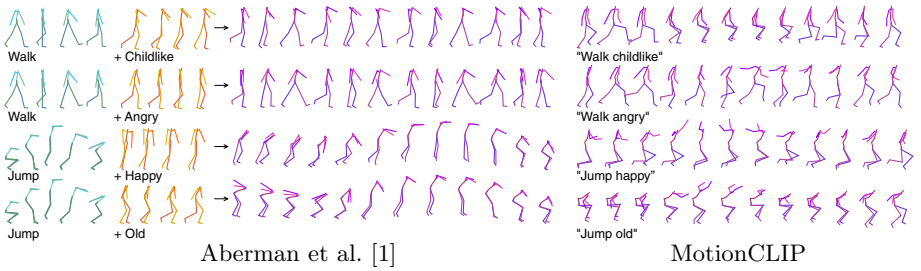

Fig. 7. Style generation. Left: style transfer by Aberman et al. [1], conditioned on action (green) and style (orange) motions. Right: MotionCLIP generating style from plain text input.

MotionCLIP variant without pretraining CLIP. Figure 5 further demonstrates that while MotionCLIP generates better motions for in-domain examples, for the out-of-domain set, it is not only the highest quality model, but often the only model that is not mode-collapsed, and generates a valid result. Figure 12 qualitatively shows the effect of text and image CLIP losses on the generation quality. In the Supplementary Materials, we present a variety of sports generated by MotionCLIP, as used in the user-study. Even though this is not a curated list, the motion created according to all 30 depicted text prompts resembles the requested actions.

Styles. We investigate MotionCLIP's ability to represent motion style, without being explicitly trained for it. We compare the results produced by MotionCLIP to the style transfer model by Aberman et al. [1]. The latter receives two input motion sequences, one indicating content and the other style, and combines them through a dedicated architecture, explicitly trained to disentangle style and content from a single sequence. In contrast, we simply feed MotionCLIP with the action and style textual names (e.g. "walk proud"). We show to

users[4] the outputs of the two models side-by-side and ask them to choose which one presents both style and/or action better (See Figure 7). Even though Aberman et al. was trained specifically for this task and gets the actual motions as an input, rather then text, Table 1 shows comparable results for the two models, with an expected favor toward Aberman et al.. This, of course, also means that MotionCLIP allows expressing style with free text, and does not require an exemplar motion to describe it. Such novel free text style augmentations are demonstrated in Figure 8.

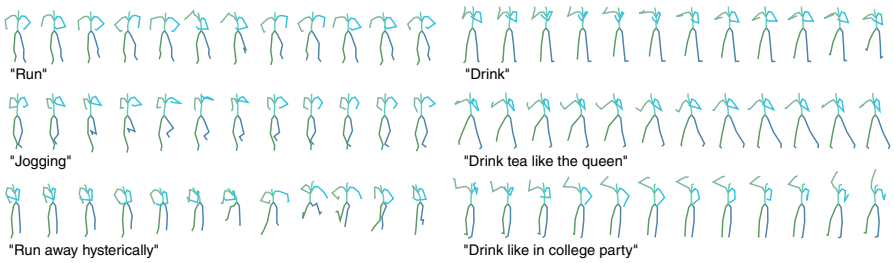

Fig. 8. MotionCLIP expresses the style described as a free text.

	Aberman et al. [1]	**MotionCLIP**
Happy	31.3%	**68.7%**
Proud	**86.4%**	13.6%
Angry	43.5%	**56.5%**
Childlike	**57.6%**	42.4%
Depressed	**74.2%**	25.8%
Drunk	**50%**	**50%**
Old	**57.7%**	42.3%
Heavy	**85.2%**	14.8%
Average	**62.1%**	37.9%

Table 1. Style generation - user study (preference score side-by-side). We compare our style + action generation from text, to those of Aberman et al. [1] which gets style and content motions as input. Interestingly, although not trained to generate style, our model wins twice and break even once

Abstract language. One of the most exciting capabilities of MotionCLIP is generating motion given text that doesn't explicitly describe motion. This includes obvious linguistic connections, such as the act of sitting down, produced from the input text "couch". Other, more surprising examples include mimicking

[4] 55 unique users, each was asked 4 questions.

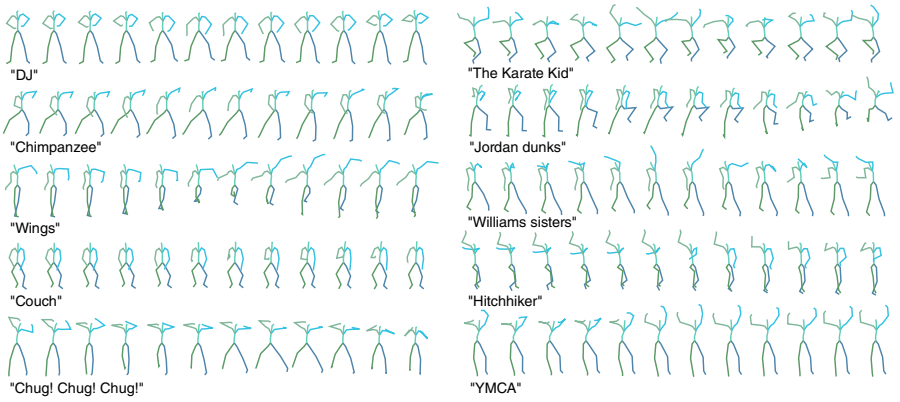

Fig. 9. Abstract language. MotionCLIP generates the signature motions of culture figures and phrases.

the signature moves of famous real and fictional figures, like *Usain Bolt* and *The Karate Kid*, and other cultural references like the famous ballet performance of *Swan Lake* and the *YMCA* dance (Figures 1 and 9). These results include motions definitely not seen during training (e.g., Spiderman in Figure 1), which strongly indicates how well the motion manifold is aligned to CLIP space.

4.3 Motion Manifold Evaluation

Accuracy metric Following ACTOR[33] we use the *accuracy* metric to evaluate MotionCLIP's latent space. To this end, we use the UESTC action recognition dataset [21], including $25K$ motion sequences annotated with 40 action classes. This data was not seen during MotionCLIP training, hence, this is a zero-shot evaluation for our model. We encode the validation set motions, and collect their mean and standard deviation. Using these statistics, we sample new motions for each class according to the class distribution found in the test set. Then, we decode the sampled motions and feed the result to an action recognition model (pre-trained on UESTC, as reported by ACTOR). In table 2, we use the accuracy metric to ablate CLIP losses, and examine 2-layered GRU [5] backbone, in addition to our reported Transformer backbone. The results imply that although failing in text-to-motion (Figure 12), GRU provides smoother latent space.

Interpolation As can be seen in Figure 10, the linear interpolation between two latent codes yields semantic transitions between motions in both time and space. This is a strong indication to the smoothness of this representation. Here, the source and target motions (top and bottom respectively) are sampled from the validation set, and between them three transitions are evenly sampled from the linear trajectory between the two representations, and decoded by MotionCLIP.

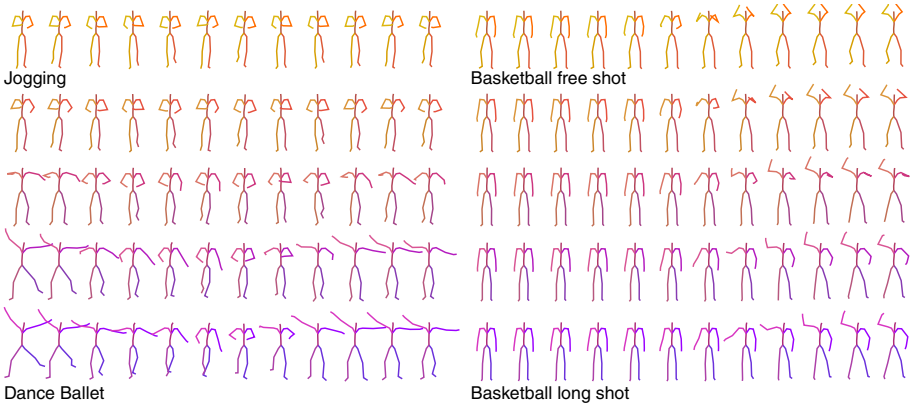

Fig. 10. Latent space motion interpolation. MotionCLIP enables semantic interpolation between two motions.

	Transformer	GRU
MotionCLIP	$63.5\% \pm 0.7$	$73.3\% \pm 0.8$
- w.o. image loss	$69\% \pm 0.8$	$62.9\% \pm 0.9$
- w.o. text loss	$63.5\% \pm 0.7$	$75.6\% \pm 0.8$
- w.o. pretraing CLIP	$45.1\% \pm 1$	$66.2\% \pm 0.8$
ACTOR [33]	$99.2\% \pm 0.1$	
Real	$98.8\% \pm 0.2$	

Table 2. Accuracy metric. We report the Top-5 accuracy of a pre-trained action recognition model [33] trained on the UESTC dataset [21] for ablated variants of MotionCLIP. This dataset was not seen during training, hence, it provides zero-shot evaluation for our latent space. The results surprisingly indicate that GRU backbone yields smoother latent space, although failing in the text-to-motion task.

4.4 Motion Manifold Applications

It is already well established that the CLIP space is smooth and expressive. We demonstrate its merits also exist in the aligned motion manifold, through the following experiments.

Latent-Based Editing To demonstrate how disentangled and uniform MotionCLIP latent space is, we experiment with latent-space arithmetic to edit motion (see Figure 11). As can be seen, these linear operations allow motion compositionality - the upper body action can be decomposed from the lower body one, and recomposed with another lower body performance. In addition, Style can be added by simply adding the vector of the style name embedding. These two properties potentially enable intuitive and semantic editing even for novice users.

Action Recognition Finally, we further demonstrate how well our latent spaces is semantically structured. We show how combined with the CLIP text

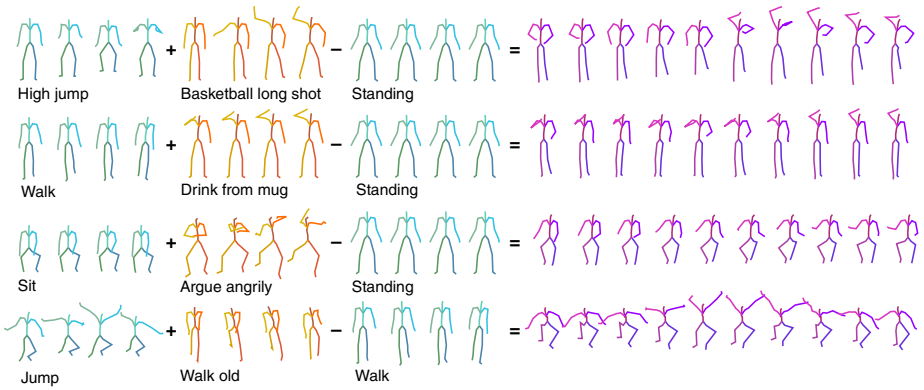

Fig. 11. Latent space motion editing. MotionCLIP enables semantic editing in latent space. Here we demonstrate two applications (1) upper and lower body action compositions (top two examples) and (2) style transfer (the two examples at the bottom).

	Top-1 acc.	Top-5 acc.
MotionCLIP	40.9 %	57.71%
- w.o. image loss	35.05%	50.26%
- w.o. text loss	4.54%	18.37%
2s-AGCN [41]	41.14%	73.18%

Table 3. Action Recognition. Using MotionCLIP together with CLIP text encoder for classification yields performance marginally close to 2s-AGCN [41] dedicated architecture on the BABEL-60 benchmark.

encoder, MotionCLIP encoder can be used for action recognition. We follow BABEL 60-classes benchmark and train the model with BABEL class names instead of the raw text. At inference, we measure the cosine distance of a given motion sequence to all 60 class name encodings and apply softmax, as suggested originally for image classification [38]. In table 3, we compare Top-1 and Top-5 accuracy of MotionCLIP classifier to 2s-AGCN classifier [41], as reported by Punnakkal et al. [37]. As can be seen, this is another example where our framework performs similarly to dedicated state-of-the-art methods, even though MotionCLIP was not designed for it.

5 Conclusions

We have presented a motion generation network that leverages the knowledge encapsulated in CLIP, allowing intuitive operations, such as text conditioned motion generation and editing. As demonstrated, training an auto-encoder on the available motion data alone struggles to generalize well, possibly due to data quality or the complexity of the domain. Nonetheless, we see that the same

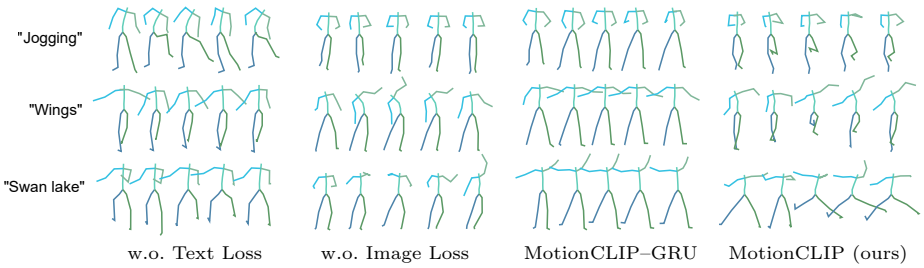

Fig. 12. Ablation study on the loss term and backbone. By training with both losses, CLIP text and CLIP image, MotionCLIP can better generate motions for challenging text inputs.

auto-encoder with the same data can lead to a significantly better understanding of the motion manifold and its semantics, merely by aligning it to a well-behaved knowledge-rich latent space. We restress the fascinating fact that even though CLIP has never seen anything from the motion domain, or any other temporal signal, its latent structure naturally induces semantics and disentanglement. This succeeds even though the connection between CLIP's latent space and the motion manifold is through sparse and inaccurate textual labeling. In essence, the alignment scheme transfers semantics by encouraging the encoder to place semantically similar samples closer together. Similarly, it induces the disentanglement built into the CLIP space, as can be seen, for example, in our latent-space arithmetic experiments. Of course, MotionCLIP has its limitations, opening several novel research opportunities. It struggles to understand directions, (e.g. left, right and counter-clockwise), to capture some styles (such as heavy and proud), and is of course not consistent for out-of-domain cultural reference exapmles (e.g, it fails to produce *Cristiano Ronaldo*'s goal celebration, and *Superman*'s signature pose). In addition, we observe that text-to-motion generation provide substandard global position and orientation, and leave it to a future work. Nonetheless, we believe MotionCLIP is an important step toward intuitive motion generation. Knowledge-rich disentangled latent spaces have already proven themselves as a flexible tool to novice users in other fields, such as facial images. In the future, we would like to further explore how powerful large-scale latent spaces could be leveraged to benefit additional domains.

Acknowledgements

We thank and appreciate the early collaboration on the problem with our colleagues Nefeli Andreou, Yiorgos Chrysanthou and Nikos Athanasiou, that fueled and motivated our research. We thank Sigal Raab, Rinon Gal and Yael Vinker for their useful suggestions and references. This research was supported in part by the Israel Science Foundation (grants no. 2492/20 and 3441/21), Len Blavatnik and the Blavatnik family foundation, and The Tel Aviv University Innovation Laboratories (TILabs).

References

1. Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., Chen, B.: Unpaired motion style transfer from video to animation. ACM Transactions on Graphics (TOG) **39**(4), 64–1 (2020)
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019)
3. Aristidou, A., Yiannakidis, A., Aberman, K., Cohen-Or, D., Shamir, A., Chrysanthou, Y.: Rhythm is a dancer: Music-driven motion synthesis with global structure. IEEE Transactions on Visualization and Computer Graphics (2022)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
6. Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648 (2016)
7. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)
8. Edwards, P., Landreth, C., Fiume, E., Singh, K.: Jali: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on graphics (TOG) **35**(4), 1–11 (2016)
9. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
10. Frans, K., Soros, L., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. arXiv preprint arXiv:2106.14843 (2021)
11. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946 (2021)
12. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
13. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
14. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
15. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 976–980. IEEE (2022)
16. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
17. He, X., Peng, Y.: Fine-grained image classification via combining vision and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5994–6002 (2017)

18. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (TOG) **35**(4), 1–11 (2016)
19. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. ACM Transactions on Graphics (TOG) **41**(4), 1–19 (2022). https://doi.org/10.1145/3528223.3530094
20. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
21. Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale rgb-d database for arbitrary-view human action recognition. In: Proceedings of the 26th ACM international Conference on Multimedia. pp. 1510–1518 (2018)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: The IEEE International Conference on Computer Vision (ICCV) (2021)
24. Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. Learning **2018**, 1 (2018)
25. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
27. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4Clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
28. Maheshwari, S., Gupta, D., Sarvadevabhatla, R.K.: Mugl: Large scale multi person conditional action generation with locomotion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 257–265 (2022)
29. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
30. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13492–13502 (2022)
31. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
33. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Vision (ICCV). pp. 10985–10995 (October 2021)
34. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022)

35. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
36. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics and Autonomous Systems **109**, 13–26 (2018)
37. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 722–731 (Jun 2021)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
39. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
40. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Fumero, M., Malekshan, K.R.: Clip-forge: Towards zero-shot text-to-shape generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18603–18613 (2022)
41. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
42. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
44. Vinker, Y., Pajouheshgar, E., Bo, J.Y., Bachmann, R.C., Bermano, A.H., Cohen-Or, D., Zamir, A., Shamir, A.: Clipasso: Semantically-aware object sketching. arXiv preprint arXiv:2202.05822 (2022)
45. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3835–3844 (2022)
46. Wang, J., Xu, H., Narasimhan, M., Wang, X.: Multi-person 3d motion prediction with multi-range transformers. Advances in Neural Information Processing Systems **34** (2021)
47. Wen, Y.H., Yang, Z., Fu, H., Gao, L., Sun, Y., Liu, Y.J.: Autoregressive stylized motion synthesis with generative flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13612–13621 (2021)
48. Yamada, T., Matsunaga, H., Ogata, T.: Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robotics and Automation Letters **3**(4), 3441–3448 (2018)
49. Youwang, K., Ji-Yeon, K., Oh, T.H.: Clip-actor: Text-driven recommendation and stylization for animating human meshes (2022)
50. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019)