

# PoseEmbroider: Towards a 3D, Visual, Semantic-aware Human Pose Representation

Ginger Delmas<sup>1,2</sup>, Philippe Weinzaepfel<sup>2</sup>  
Francesc Moreno-Noguer<sup>1</sup>, and Grégory Rogez<sup>2</sup>

<sup>1</sup> Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain

<sup>2</sup> NAVER LABS Europe

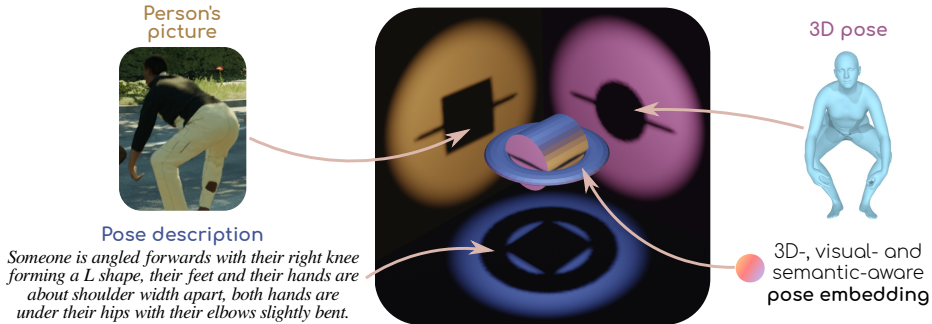
<https://europe.naverlabs.com/research/PoseEmbroider/>

**Abstract.** Aligning multiple modalities in a latent space, such as images and texts, has shown to produce powerful semantic visual representations, fueling tasks like image captioning, text-to-image generation, or image grounding. In the context of human-centric vision, albeit CLIP-like representations encode most standard human poses relatively well (such as standing or sitting), they lack sufficient acuteness to discern detailed or uncommon ones. Actually, while 3D human poses have been often associated with images (*e.g.* to perform pose estimation or pose-conditioned image generation), or more recently with text (*e.g.* for text-to-pose generation), they have seldom been paired with both. In this work, we combine 3D poses, person’s pictures and textual pose descriptions to produce an enhanced 3D-, visual- and semantic-aware human pose representation. We introduce a new transformer-based model, trained in a retrieval fashion, which can take as input any combination of the aforementioned modalities. When composing modalities, it outperforms a standard multi-modal alignment retrieval model, making it possible to sort out partial information (*e.g.* image with the lower body occluded). We showcase the potential of such an embroidered pose representation for (1) SMPL regression from image with optional text cue; and (2) on the task of fine-grained instruction generation, which consists in generating a text that describes how to move from one 3D pose to another (as a fitness coach). Unlike prior works, our model can take any kind of input (image and/or pose) without retraining.

**Keywords:** 3D Human Pose · Multi-Modal Retrieval · Text Generation

## 1 Introduction

People play a central role in many applications across a wide range of domains, including robotics, digitization (such as virtual avatars), and entertainment. In many of these contexts, the human pose is a defining characteristic. While a large body of work aims to estimate [8, 27, 42] or predict it [1, 6, 50], for instance, to further facilitate human-robot interaction, another seeks to generate it [24, 28, 38, 58, 60], to enhance experiences in video games or virtual worlds. These tasks demonstrate the crucial importance of human understanding.



**Fig. 1: Motivation.** Comprehending a complex 3D object in a 2D world is not simple. Having access to several of its shadows, obtained by lighting it under different angles, can help better understand it. Similarly, we collect several multi-modal (and naturally partial) observations of the human pose (the “shadows”), and try to create an enriched pose embedding (the “3D” object). This embedding is derived from 3D joint rotations, pictures of humans and pose descriptions, then further used in downstream applications requiring human pose understanding.

Early works have focused on detecting and visually understanding people. While human bodies can already be fairly well studied through visual data, true human understanding goes beyond mere perception. It also relies on *meaning*, that is, semantics. Now, we humans, tend to prefer when the world’s semantics match ours. This is where natural language comes into play. Language empowers the conveyance of complex and abstract concepts; making it possible to gather similar elements together under the same word. For instance, one person could have their hand at shoulder level, and another person their hand way overhead; yet, both individuals could be “*waving*”.

Ultimately, both visual and textual data are essential to achieve human understanding: they are two facets of the same prism. However, both are imperfect: visual data may exhibit occlusions or depth uncertainty, while text is relatively ambiguous. Despite these flaws, they provide crucial information that a 3D pose alone could not convey, such as world affordance, reality anchoring, and semantics. In the end, all three modalities (visual data, text and 3D poses) can be considered complementary – partial, yet valuable – observations of the same abstract “human pose” concept (see Figure 1 for an illustration).

More concretely, recent advances have demonstrated the utility of pairing images and texts to derive powerful semantic image embeddings [61]. In this work, we extend this principle to the concept of human pose. We aim to derive a rich pose embedding that is simultaneously semantic-, visual- and 3D-aware, by embroidering images, texts and 3D poses together. Indeed, current endeavors only yield coarse representations of human poses, failing to distinguish between two similar complex poses.

Previous works have essentially focused on connecting individuals depicted in images to their 2D or 3D pose [12, 27, 71, 74], or on linking 3D poses with fine-grained text descriptions [15], thereby producing strong visual pose embeddings

or semantic pose embeddings. More recently [21] repurposed a large vision and language transformer model to output a human pose, based on either an image, text or a combination of both inputs. However, just like other multi-modal works leveraging large language models (LLM) [19, 70], it requires converting new modalities to equivalent textual representations, so as to enable processing in the LLM space. This process could lead to partial loss of modality-specific information, in particular information that cannot be transcribed through text. Another body of works [14, 64, 69] proposes to unify human-centric perception tasks under a single multi-modal model. Yet, these models are generally trained with task-specific objectives. Overall, recent multi-modal methods tend to *align* modalities to enable any-to-any translation [25, 52], but do not necessarily combine multi-modal information to build a single versatile representation.

In this paper, we design a multi-modal framework that *embroiders* different modalities, so as to build a richer semantic-, visual-, 3D-aware pose embedding space. We use a transformer to aggregate information from available modalities within a single global token. The model is trained with uni-modal contrastive objectives, on the reprojections of this global representation to each modality space. As a result, we can enhance any single modality embedding fed to our model with multi-modal awareness. We demonstrate the benefit of our proposed pose representation by addressing the tasks of any-to-any multi-modal retrieval, pose estimation, as well as pose instruction generation, which has a direct application in automatic fitness coaching. This task consists in producing a text that specifies how to modify one pose into another. Differing from the initially proposed baseline PoseFix [16], the utilization of our multi-modal representation makes it possible to process direct camera input without the need for additional retraining. In summary, our contributions list as follow:

- \* We introduce a new framework to embroider together several human pose-related modalities and derive a rich semantic-, visual-, 3D-aware pose embedding space (Section 3). We train it on the adapted BEDLAM-Script dataset (a description-augmented version of BEDLAM [7]). As a direct side-product, we present results for any-to-any multi-modal retrieval (Section 4).
- \* We showcase an application of the proposed enhanced pose representation for the task of pose instruction generation (Section 5). Although our method is almost exclusively trained on synthetic data (the proposed BEDLAM-Fix dataset), we obtain promising results on real-world images.
- \* We illustrate SMPL regression as another application (Section 6).

## 2 Related Work

We propose a novel *multi-modal human pose representation*, using a framework related to general *multi-modal alignment*, which can be applied to downstream tasks such as *pose instruction generation*. We briefly review related methods.

**Multi-modal representations of humans.** Several methods proposed to learn efficient pose-structured human image representations [11, 12, 74], but do not consider valuable extraneous information (3D, or finer-semantics brought by

text data). A growing body of work, focusing on human-centric perception, use more human-related multi-modal data (*e.g.* RGB images, depth maps, 2D key-points, 3D pose and shape, visual attributes or description *etc.*) [32] to perform diverse tasks, *e.g.* person re-identification, human parsing, pose estimation, action recognition, attribute recognition *etc.* These works usually resort to task- or dataset-specific objective functions, to train a single model taking multi-modal input. PATH [64] learns a shared image transformer with task-specialized projectors and dataset-dedicated heads. UniHCP [14] processes the image via cross-attentions taking task-specific queries, task-wise interpreted as a set of features directly denoting the expected task output. Less visual-centric, Hulk [69] trains modality-specific tokenizers which outputs are processed in a shared encoder and decoded based on given modality indicators (“query tokens”) to perform modality translation. Unlike these works, we introduce a task-agnostic multi-modal-aware representation, that could be used out-of-the box in any pose-related task. In particular, we show its effectiveness on the task of pose instruction generation from images, which requires adequate human pose perception and a fine-grained semantic understanding of the body/ parts and their relationships.

Closer to this idea, PoseScript [15] models semantic pose embeddings by pairing 3D poses and descriptions in natural language. However it does not consider the visual modality. More recently, ChatPose [21] append a SMPL [49] projection layer to a large vision and language model [48], so to leverage its reasoning abilities for pose estimation and text-to-pose generation. They hence derive a visual-semantic 3D pose representation, yet constrained to live in the textual space. Also, the model is not designed to take direct 3D pose input.

**Multi-modal alignment.** It is common to align different modalities to perform multi-modal applications. Efforts spanned aligning text and images [20, 23, 41, 43], videos [2, 3, 47], audio [33, 54], robotic states [19], 3D shapes [62], 3D scenes [34], 3D human poses [15], human motions [59, 72], and so forth. Beyond empowering cross-modal retrieval, connecting modalities gives birth to powerful multi-informed versatile encodings. One of the most recent iconic works is CLIP [61], which learns a joint embedding space for images and texts with contrastive learning. The produced visual semantic representations are reused off-the-shelf in a variety of tasks and domains [17, 36, 65, 73]. Research in multi-modal alignment further stepped up thanks to the introduction of ever-growing datasets, computational resources and models, which make it possible to get qualitative pseudo labeling [52, 71] or reliable synthetic data [9].

Some works have explored aligning more than two modalities. Omnivore [26] aligns several unpaired labeled visual modalities by feeding all visual patches to a single transformer, trained for classification. All modalities end up being encoded in the same space, hatching cross-modal retrieval. ImageBind [25] brings it one step further by additionally considering non-visual modalities such as audio and text, leveraging the natural co-occurrence of images with other modalities.

Other recent works align modalities to facilitate any-to-any generation. NExT-GPT [70], similar to Palm-e [19], feeds pretrained uni-modal representations to learnable modality-specific projections layers, such that they can be processed



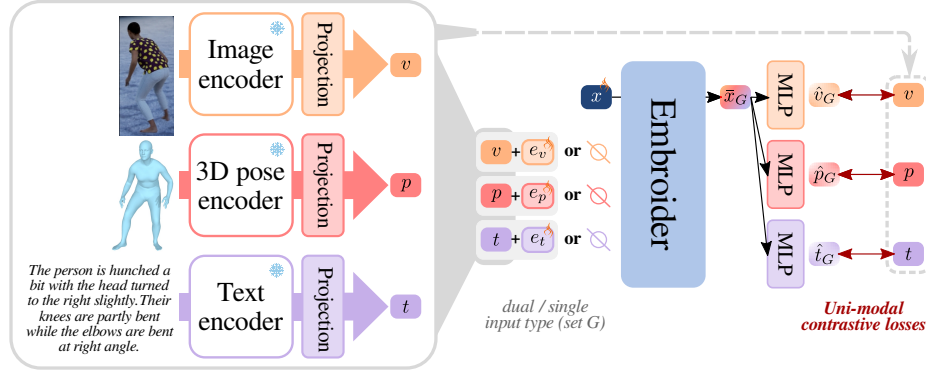
by a frozen LLM [13]. Eventually, the outputs are re-projected in uni-modal spaces and fed to pretrained diffusion models for generation. 4M [52] tokenizes all modalities to process them with the same transformer decoder, trained with masked modeling (for visual modalities) and next token prediction (for sequence modalities), for a random subset of (modality) query and target tokens. It converts modalities but does not learn a single multi-modal-informed representation.

Different from the above-mentioned methods, we go beyond *aligning* or *converting* modalities. We learn to *intermingle* them with the aim of obtaining a unique, richer, multi-modal-informed representation, computable from any set of input modalities. Basically, we use multi-modal data to figure how to *enhance* uni-modal encodings. In particular, this augmented embedding is not compelled to live in the textual space as in [19, 21, 70], which could lead to the loss of non-textual information. Instead, it is free to assume any relevant structure in its own embedding space. For the same reason, we apply contrastive learning on the modality-specific spaces instead of the augmented embedding space, using uni-modal reprojections of the augmented encoding.

**Pose instruction generation** is a recent task, which consists in generating an instruction explaining how to correct one pose in another specific pose. FixMyPose [37] introduced a first dataset based on highly-synthetic pairs of images. AIFit [22] focuses on video data, and learns to produce feedbacks out of template sentences, based on the comparison between a trainee’s and a trainer’s motion extracted features. More recently, PoseFix [16] adapted the automatic captioning pipeline from [15] to create synthetic instructions for a pair of 3D poses sampled from AMASS [51]. Those prove useful for pretraining, before finetuning on a small set of human-written texts. The proposed text generation model is a simple text decoder conditioned on pose pairs via cross-attentions. However, it is limited to parameterized pose input (*i.e.* 3D joint rotations), and thus cannot be directly applied to real-world scenarios, as in a fitness coaching application receiving camera input. In this work, we use our 3D-, visual- and semantic-aware embedding to scale the task on direct image input, without having to train the model on a dataset of real images and textual instructions together.

### 3 The PoseEmbroider framework

We now describe our proposed framework for learning multi-modal enhanced pose representations, see Figure 2 for an illustration. Note that the overall design does not rely on specific types or numbers of modalities, allowing for its extension to other domains and sets of modalities. In this paper, we focus on three modalities: images of people, 3D human poses (parameterized by the rotations of the main SMPL [49] body joints) and text, in the form of fine-grained pose descriptions in natural language. Each of them provide different kinds of information, be it visual, spatial and kinematic, or semantic. We aim to leverage their partial representation of the same abstract concept of human pose, to build a richer pose embedding. For simplicity, we assume in what follows that we have a tri-modal dataset, *i.e.* with samples from all modalities for each example.



**Fig. 2: The PoseEmbroider framework.** Each modality is encoded independently by an encoder (left). The PoseEmbroider (right) is a transformer-based model, taking a varying set of modality inputs. It produces a visual-, 3D-, semantic-aware pose representation  $\bar{x}$ , by embroidering together available inputs. The model is trained using uni-modal contrastive losses between the modality-specific reprojections  $\hat{m} \in \{\hat{v}, \hat{p}, \hat{t}\}$  of  $\bar{x}$  and the original modality encodings  $m \in \{v, p, t\}$ . The total objective function accounts for various  $\bar{x}_G$ , obtained from the set  $G$  of input modalities.  $x$  and  $e_m$  are learnable tokens, ‘+’ denotes an addition.

### 3.1 Method

**Encoders.** Similar to other multi-modal methods [25], we resort to pretrained uni-modal encoders for each modality. Specifically, we use a Vision Transformer [18] tuned on human data [57] to encode images; a variant of the VPoser [57] encoder for poses (trained on the main 22 body joints); and a text transformer [66] mounted on top of DistilBERT [63] frozen word embeddings, obtained from a text-to-pose retrieval model [15]. All encoders are used frozen.

**General framework.** Each modality input is first processed by its respective frozen pretrained encoder, then fed to a modality-specific learnable linear layer followed by a ReLU activation, to select pose-related features and filter out irrelevant details (*e.g.* background information in images). Average pooling further reduces multi-token representations into single vectors (image case). Let  $v$ ,  $p$  and  $t$  in  $\mathbb{R}^d$  denote the corresponding outputs for the image, pose and text of a data triplet respectively. In what follows, we use  $m$  to refer to any of them:  $m \in M := \{v, p, t\}$ .

The *PoseEmbroider* mainly consists of a transformer [66]. It takes a variable set of input modalities  $G \in S$ , in addition of a learnable global token  $x$ , that will collect and aggregate pose knowledge across all input modalities through the attention mechanisms of the transformer. We consider any combination of one (“single input” type) or two (“dual input” type) input modalities, *i.e.*  $S := \{\{v\}, \{p\}, \{t\}, \{v, p\}, \{v, t\}, \{p, t\}\}$ .

Hence, the PoseEmbroider is provided with  $\{x\} \cup G$ , where a modality-specific learnable token  $e_m \in \mathbb{R}^d$  has been added to each input modality encoding in

order to inform the transformer about their nature.  $e_m$  can be compared to a special kind of learnable positional encoding.

The PoseEmbroider outputs  $|G| + 1$  tokens. Yet, we only consider the first one, noted  $\bar{x}_G$ , which derives directly from the token  $x$  and holds specific information from  $G$ . It represents the richer, multi-modal informed pose embedding, illustrated as the 3D object in Fig. 1. It can be obtained from any set of input modalities, and be used as main pose representation in downstream tasks.

**Training.** To ensure  $\bar{x}_G$  carries important visual, spatial & kinematic and semantic pose information, we compare it to each of the original unimodal encodings. However, we do not perform a direct comparison of  $\bar{x}_G$  with each modality encoding  $m \in M$ , as it would compel all modalities to live in the same space, and eventually lead to the collapse of  $\bar{x}_G$  to a representation of *common* information between modalities. Instead, we want  $\bar{x}_G$  to be an *enhancement* of its components. Even more, we want it to form sensible postulates for the modalities that did not directly contributed to its derivation.

Thus, to train  $\bar{x}_G$ , we project it “back” to each modality space thanks to expendable modality-specific multi-layer perceptrons [30] (MLPs). These yield  $\hat{m}_G \in \hat{M}_G := \{\hat{v}_G, \hat{p}_G, \hat{t}_G\}$ . For a given batch of  $B$  training samples, we then compute the uni-modal contrastive loss for each modality  $m$ , following the widely used InfoNCE [55]:

$$\mathcal{L}_c(y, z) = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\gamma \sigma(y_i, z_i))}{\sum_j \exp(\gamma \sigma(y_i, z_j))}, \quad (1)$$

where  $\gamma$  is a learnable temperature parameter and  $\sigma$  is the cosine similarity function defined as  $\sigma(y, z) = y^\top z / (\|y\| \|z\|)$ . Denoting  $\mathcal{M}_G := \{(m, \hat{m}_G) \mid m \in M, \hat{m}_G \in \hat{M}_G \text{ of the same modality}\}$ , the total loss is then:

$$\mathcal{L} = \sum_{G \in S} \sum_{\mathcal{M}_G} \mathcal{L}_c(m, \hat{m}_G). \quad (2)$$

Metaphorically, if we refer to Figure 1, we used available shadows ( $G$ ) to try to infer the 3D object ( $\bar{x}_G$ ), thanks to the PoseEmbroider. To optimize the latter, we light the object under different angles to check shadow consistency in a soft way ( $\mathcal{L}$ ). Specifically, we do not require the shadows to perfectly match (as it would be the case with a reconstruction loss): we only enforce the ranking of the real object’s shadow to be better than another object’s shadow. Actually, during this “validation” step, we assume access to all ground-truth shadows: even if one or more modalities were missing from the input, as *e.g.* with  $\{p\}$ , the loss is applied on *all* available modalities. This design forces  $\bar{x}_G$  to be multi-modal aware, beyond being simply multi-modal informed. In other words, the PoseEmbroider aims at providing a strong representation of any (partial) combination of the modalities.

### 3.2 Dataset: BEDLAM-Script

A multi-modal model requires multi-modal data. However, there is no existing dataset that gathers images, 3D poses and texts all at once. In fact, perfect

3D pose labeling generally requires expensive and non-scalable in-studio capture. Therefore, datasets with real-world images and good-quality 3D human pose annotations are rare, which motivates the creation of synthetic datasets. BEDLAM [7] is the most recent endeavor in this regard. It provides rendered sequences of clothed humans in different environments, performing a wide variety of motions extracted from the AMASS dataset [51]. It thus comes with ground-truth 3D pose annotations. Previous works [10] have shown that training on BEDLAM brought the best result for pose estimation over various real data benchmarks [46, 68, 75], compared to training on other (including real) datasets.

We thus opt for this dataset. Similar to [15], we first select a set of  $N$  diverse poses by farthest point sampling, *i.e.* sampling iteratively the pose that has the largest mean-per-joint distance with respect to the set of poses already selected. This process makes it possible to efficiently reduce the size of the training ( $N=50k$ ) and validation ( $N=10k$ ) sets while preserving data diversity. We augment each image-pose pair with 3 detailed pose descriptions using the automatic captioning pipeline from PoseScript [15]. Specifically, given 3D joint coordinates, they compute a collection of “posecodes” informing about atomic pose configurations (*e.g.* bending of a body part, relative body part positioning, *etc.*). Those are further converted to natural language description thanks to a set of syntactic rules, merging posecodes that carry similar semantic information. We improve this pipeline to account for head rotations and self-contacts, so as to get better pose descriptions. We do so using a mesh rendering of the pose, and a self-contact detection algorithm [53] coupled with a semantic segmentation of body vertices. We refer to the resulting dataset of images, 3D poses and text descriptions as BEDLAM-Script, and train our PoseEmbroider framework on it. While its training involves exclusively synthetic data, we show that the PoseEmbroider produces convincing results on real-world images and human-written texts.

**Data processing.** We consider normalized 3D body poses, *i.e.* with the global rotation set such that the hips are aligned and always facing in the same direction. The motivation is to force the model to extract more general, world-anchored pose knowledge, in contrast to camera-dependent pose information. While BEDLAM annotations are in SMPL-X [57] format, *i.e.* they include hands, we restrict the 3D pose representation to the main 22 joints of the body. Future work could additionally consider the hands, by also adapting the automatic captioning pipeline to provide such information, *e.g.* as in [45].

## 4 Results on Multi-Modal Retrieval

As a direct side product of its training, the PoseEmbroider framework exhibits multi-modal retrieval abilities. In this section, we report results for any-to-any multi-modal retrieval, and use this task for our ablations. We additionally show-case qualitative results for edited-retrieval in a multi-modal setting.

**Evaluation metrics.** We consider all possible any-to-any multi-modal retrieval sub-tasks. This results in 6 single-query and 3 dual-query tasks. The standard metric for retrieval evaluation is the recall@K ( $R@K$ ), *i.e.* the percentage of

**Table 1: Multi-modal retrieval results.** Models are trained on BEDLAM-Script and evaluated on its validation set. The *total* mRecall is the average of *single* and *dual*, corresponding to the average over all single- and dual-query retrieval tasks respectively. V, P, and T refer to the “visual” (image), “pose” and “text” modalities respectively. The aligner trained on single-input only (first row) corresponds to the idea of [15, 25].

|                                                    | mRecall↑     |               |             | Single query |      |      |      |      |      | Dual query |      |      |
|----------------------------------------------------|--------------|---------------|-------------|--------------|------|------|------|------|------|------------|------|------|
|                                                    | <i>total</i> | <i>single</i> | <i>dual</i> | V→P          | V→T  | P→V  | P→T  | T→V  | T→P  | VP→T       | PT→V | VT→P |
| <b>Representation &amp; training input subsets</b> |              |               |             |              |      |      |      |      |      |            |      |      |
| Aligner (single-input only)                        | 72.4         | 66.5          | 78.3        | 77.5         | 46.3 | 75.8 | 76.0 | 46.2 | 77.5 | 71.3       | 70.7 | 92.8 |
| Aligner (dual-input extension)                     | 72.5         | 66.4          | 78.5        | 76.9         | 45.8 | 75.8 | 76.3 | 45.9 | 77.8 | 72.0       | 70.4 | 93.1 |
| PoseEmbroider (single input only)                  | 69.7         | 66.7          | 72.7        | 80.2         | 48.0 | 74.6 | 77.7 | 43.6 | 76.4 | 67.5       | 61.9 | 88.7 |
| PoseEmbroider (dual input only)                    | 71.1         | 58.7          | 83.6        | 69.7         | 30.0 | 78.0 | 78.7 | 26.5 | 69.2 | 79.6       | 78.2 | 93.0 |
| PoseEmbroider (S)                                  | 74.6         | 66.9          | 82.2        | 79.7         | 47.8 | 76.2 | 78.7 | 43.2 | 75.8 | 77.9       | 75.1 | 93.7 |
| <b>PoseEmbroider architecture</b>                  |              |               |             |              |      |      |      |      |      |            |      |      |
| MLP core                                           | 73.4         | 66.5          | 80.3        | 79.9         | 46.4 | 75.5 | 79.2 | 41.9 | 76.2 | 76.8       | 71.1 | 93.0 |
| Trans. core, no projection heads                   | 73.5         | 66.4          | 80.5        | 80.2         | 46.7 | 75.9 | 76.5 | 43.6 | 75.6 | 75.0       | 73.6 | 93.0 |
| Transf. core, w/ projection heads (proposed)       | 74.6         | 66.9          | 82.2        | 79.7         | 47.8 | 76.2 | 78.7 | 43.2 | 75.8 | 77.9       | 75.1 | 93.7 |

queries whose annotated target appears within the top- $K$  of retrieved elements. We report the average recall over  $K \in \{1, 5, 10\}$ .

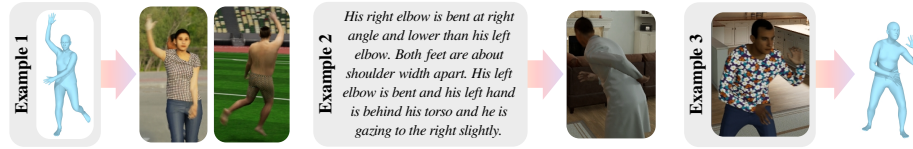
**Alignment baselines.** To highlight the benefits of the PoseEmbroider representation over a more typical alignment-based representation, we further introduce the Aligner model (similar in number of learnable parameters to the full PoseEmbroider). Unlike the PoseEmbroider, the frozen pretrained uni-modal encoders in the Aligner are followed by *deep* learnable modality-specific projection heads (*i.e.* MLPs, as opposed to single modality-specific layers leading to a shared transformer). The MLP heads are trained with pair-wise and triplet-wise alignment losses to produce a joint embedding space:

$$\mathcal{L} = \sum_{G \in S} \sum_{m \notin G} \mathcal{L}_c \left( m, \frac{1}{|G|} \sum_{q \in G} q \right). \quad (3)$$

Simply put, there is one contrastive loss term  $\mathcal{L}_c(m_1, m_2)$  for each pair of modalities (single input), and one for each kind of dual input, where the dual query representation is computed as the average of its components’ features. We denote as *Aligner (single-input only)* the model trained solely on  $S' := \{\{v\}, \{p\}, \{t\}\}$ , and as *Aligner (dual-input extension)* the model trained on  $S$ .

The *Aligner (single-input only)* can be thought as a version of ImageBind [25] applied to the human pose domain, or as a version of the PoseScript retrieval model [15] connected with an image network. Yet, different from these approaches, and *to allow a fair comparison* with the PoseEmbroider, the core encoders are not optimized: solely the MLP heads are (and are trained on BEDLAM-Script as well). Eventually, the *Aligner (dual-input extension)* explicitly integrates compositionality in the training objective, conversely to [25].

**Quantitative results** are presented in Table 1. First of all, they reveal that our proposed PoseEmbroider (row 5) outperforms the best Aligner baseline (row 2) by 2.9% (mRecall), showing particular progress with respect to dual queries (4.7%). It suggests that the PoseEmbroider not only enhances single-modality encodings but also effectively combines their knowledge. It is especially blatant



**Fig. 3: Qualitative examples of any-to-any multi-modal retrieval** on the validation split of BEDLAM-Script, for diverse input and output modalities.

for pose retrieval, where the use of both image and text as input improves over using each alone (+17.6% and +23.6%, respectively). Other cases (*e.g.* image retrieval) hint at the PoseEmbroider ability to extract intel from the most informative modality. Note that results involving both the visual and textual modalities are the lowest for all models because they are the most ambiguous (occlusions, truncations by image boundaries, incomplete/imprecise descriptions).

**Query set ablation.** It stems that the Aligner design is not very sensitive to an enhanced optimization using input combinations (row 2 *vs.* row 1). In contrast, it appears clearly that training on various sets of inputs ( $S$ ) is valuable to the PoseEmbroider, as it improves the mean Recall by +7% compared to using single inputs only (row 5 *vs.* row 3). Unsurprisingly, we also observe better performance for single queries when considering single input types during training (+13.6%), and likewise for dual queries (+15.0%, row 3 *vs.* row 4). Eventually, the best performance for all retrieval tasks is reached when using both query types.

**Architecture ablation.** To justify the PoseEmbroider design, we first replace the transformer with an MLP (row 6 in Table 1). In this setting, all modality encodings at stake are added together and fed to an MLP, whose output plays the role of  $\bar{x}_G$ . Unlike concatenation, the addition operation allows the model to be run on various subsets of modalities. This model is trained with the same objective as the proposed PoseEmbroider. Next, we ablate the re-projection heads which make it possible to use uni-modal contrastive losses. This configuration has a training objective similar to a regular alignment model. Results reveal that the transformer version is slightly more powerful than the MLP version (+1.6%), and that the re-projection heads are valuable (+1.5%), especially in the dual-query case (+2.1%).

**Qualitative results.** Figure 3 presents some results for any-to-any multi-modal retrieval, demonstrating that the PoseEmbroider efficiently associates the different pose modalities and exhibits human pose understanding.

We further consider the special use case of “edited-retrieval” where, for instance, the user is looking for a 3D pose similar to the one depicted in an image, yet a little different. The parts of the image showing unwanted pose traits are masked while supplementary information is provided through text input. We observe that the model is able to leverage and combine information from both modalities to find relevant poses. Examples are shown in Figure 4.



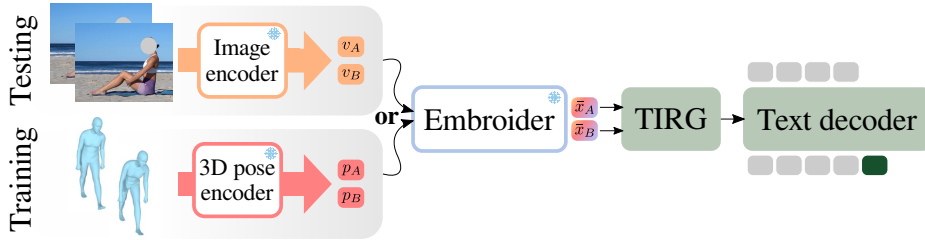
**Fig. 4: Qualitative examples of edited-retrieval in a multi-modal setting on BEDLAM-Script.** Texts specify new traits with respect to the original pose shown in the image. Artificial occlusion is created by overlaying a black rectangle on the image.

## 5 Results on human pose instruction generation

Human pose instruction generation [16, 22, 37] consists in generating directions in natural language to correct a human pose. This task has direct application in at-home fitness coaching, to provide automatic feedback. It can be solved with a text decoder, conditioned on both the *source* pose  $A$  (the trainee’s) and the *target* pose  $B$  (the trainer’s). The poses could be highly similar, and differ only in subtle aspects. Hence, this task typically requires a fine-grained semantic understanding of the human pose. Previous works [16] have proposed methods that operate on 3D pose inputs, however they cannot handle real-world scenarios where the user simply works out in front of their phone camera. To further evaluate our proposed PoseEmbroider, we replace the original 3D pose encoder in [16] by our pretrained PoseEmbroider. This configuration makes it possible to train the text generation model on reliable 3D poses, and seamlessly transfer to visual inputs, without requiring further training.

**Datasets: BEDLAM-Fix, PoseFix-OOS.** Available datasets for this task include PoseFix [16] and FixMyPose [37]. Both have approximately the same training size (4-6k), and provide human-written annotations, however the first one pairs 3D human poses from a wide variety of AMASS [51] sequences, while the second one links highly synthetic images of poses extracted from about 20 Unity dance motions. We resort to PoseFix for finetuning the text decoder, and restrict to out-of-sequence (OOS) pairs to eliminate noise stemming from global rotation changes. Similar to BEDLAM-Script (Section 3.2), we process BEDLAM [7] to create pretraining data, following the same procedure as in [16]. Specifically, we sample pose pairs from BEDLAM-Script by enforcing both semantic similarity and minimal pose difference constraints. We consider both pairs of poses performed by the same subject (*i.e.* with the same appearance, environment and motion) and different subjects. We further run their automatic *comparative* pipeline on the 3D poses to obtain synthetic instruction texts. We end up with 54k (resp. 12k) training (resp. validation) pairs.

**Method overview.** We use a similar method as [16]: the elements  $A$  and  $B$  are encoded through siamese networks, and fused thanks to the TIRG module [67] to condition an auto-regressive text decoder. We experiment with both the *Aligner* (*dual-input extension*) and the PoseEmbroider for encoding the inputs. To train, we use their cached (frozen) pretrained features, obtained from the 3D poses



**Fig. 5: The pose instruction generation model.** We train the model on pairs of poses  $(p_A, p_B)$  and use our frozen PoseEmbroider to encode them. These two embeddings are fused with TIRG [67], whose output is used to condition an auto-regressive transformer text decoder via cross-attentions. At test time, the trained model can be directly applied on poses, images or a mix of both.

**Table 2: Text generation results for different query types.** Models are trained on BEDLAM-Fix using pairs of poses only, and evaluated on the associated validation split for queries of different natures. We further finetune the text decoder on a mix of BEDLAM-Fix and PoseFix-OOS data, and report results on the test set of PoseFix-OOS. The Aligner baseline represents [16].

| Dataset (query type)                | Representation       | R Precision $\uparrow$ |             |             | NLP $\uparrow$ |             |             |
|-------------------------------------|----------------------|------------------------|-------------|-------------|----------------|-------------|-------------|
|                                     |                      | R@1                    | R@2         | R@3         | BLEU-4         | ROUGE-L     | METEOR      |
| BEDLAM-Fix $(P_A, P_B)$             | Aligner              | 31.7                   | 42.4        | 49.4        | 32.9           | 42.2        | 40.1        |
|                                     | <b>PoseEmbroider</b> | <b>43.1</b>            | <b>55.1</b> | <b>62.0</b> | <b>33.0</b>    | <b>42.6</b> | <b>40.7</b> |
| BEDLAM-Fix $(V_A, V_B)$             | Aligner              | 12.3                   | 19.2        | 23.7        | 27.8           | 37.8        | 38.9        |
|                                     | <b>PoseEmbroider</b> | <b>15.6</b>            | <b>22.6</b> | <b>27.1</b> | 27.4           | 37.2        | 37.3        |
| BEDLAM-Fix $(P_A, V_B)$             | Aligner              | 16.6                   | 24.2        | 29.8        | 28.9           | 38.7        | 40.4        |
|                                     | <b>PoseEmbroider</b> | <b>21.2</b>            | <b>29.6</b> | <b>35.6</b> | <b>29.8</b>    | 38.3        | 38.6        |
| BEDLAM-Fix $(V_A, P_B)$             | Aligner              | 21.0                   | 29.2        | 34.9        | 30.3           | 39.1        | 39.7        |
|                                     | <b>PoseEmbroider</b> | <b>29.2</b>            | <b>38.3</b> | <b>44.2</b> | <b>30.6</b>    | <b>39.3</b> | <b>40.4</b> |
| BEDLAM-Fix $(V_A + P_A, V_B + P_B)$ | Aligner              | 26.5                   | 36.4        | 42.4        | 33.1           | 41.0        | 40.1        |
|                                     | <b>PoseEmbroider</b> | <b>33.8</b>            | <b>45.0</b> | <b>51.9</b> | 31.3           | <b>41.3</b> | 39.5        |
| PoseFix-OOS $(P_A, P_B)$            | Aligner              | 19.5                   | 27.6        | 34.8        | 8.0            | 27.0        | 28.0        |
|                                     | <b>PoseEmbroider</b> | <b>27.4</b>            | <b>36.8</b> | <b>42.9</b> | <b>10.2</b>    | <b>28.8</b> | <b>29.3</b> |

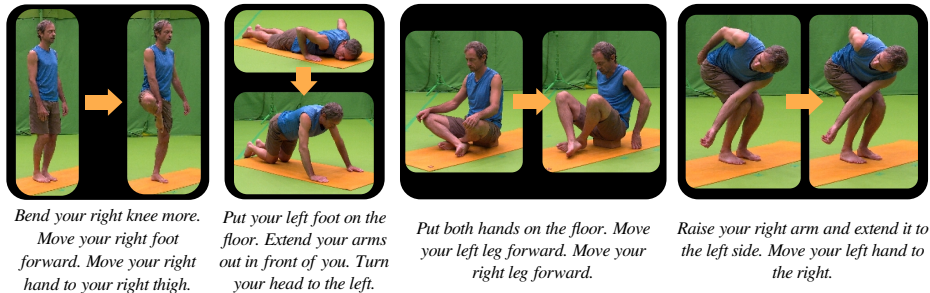
data. Sole the fusing module and the text decoder are being learned. At inference time, we use any combination of 3D poses and visual input, see Figure 5.

**Baseline.** We use our best Aligner to represent the PoseFix baseline [16]. Indeed, their encoder takes only 3D pose input, and is trained specifically for the task, alongside the text decoder. Yet, we aim to compare off-the-shelf representations, and further empower inference from visual input.

**Evaluation metrics.** Following [16], we retrain an instruction-to-pair retrieval model on the aforementioned datasets to assess the semantic content of the generated text through R-precision (with a larger, harder pool of 200). This complements the typical n-gram based NLP metrics (BLEU-4 [56], Rouge-L [44] and METEOR [5]) which evaluate formulation similarity with the reference text.

**Quantitative results** are reported in Table 2, for different sets of inputs, depending on the respective nature of elements  $A$  and  $B$  at test time (*i.e.* 3D pose





**Fig. 6: Instruction generations on real-world images using the PoseEmbroider pose representation.** The text model was trained using the frozen PoseEmbroider embeddings of 3D poses only. The generated text is shown below each image pairs.

or image). The PoseEmbroider representation outperforms the Aligner representation in all cases, particularly when both inputs are 3D poses (+36% R@1 on BEDLAM-Fix, +41% on PoseFix-OOS), despite sharing the same 3D pose encoder at the core. This suggests that the PoseEmbroider is indeed capable of *enhancing* semantic pose representations. Notably, 3D pose inputs yield better results than visual inputs, which are inherently less reliable (occlusions). Since instruction generation is driven by element  $B$ , it makes sense to find better results for the setting  $(V_A, P_B)$  than  $(P_A, V_B)$ , when compared to  $(V_A, V_B)$  (+87% *vs.* +36%). The setting  $(V_A, P_B)$  typically corresponds to that of a fitness application scenario, involving camera input from the user and clean, 3D pose registrations of the target pose.

**Qualitative results.** In Figure 6, we present examples of generated instructions for real-world input images, illustrating the steps for performing Yoga poses. While the text generation model was trained using only a dataset of 3D poses and texts, the PoseEmbroider makes it possible to transfer to image input.

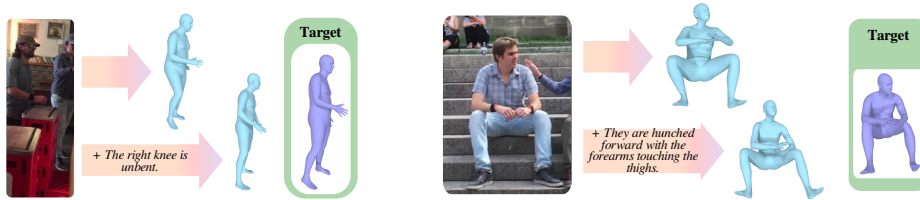
## 6 Results on SMPL regression

We showcase results for the task of SMPL regression, where the goal is to predict the pose and shape parameters of the SMPL body model [49] for a given input data of any modality. This task is known as 3D Human Mesh Recovery [35], when applied on images. We proceed similarly as before, and train a neural head to predict SMPL parameters from pretrained, frozen features of the PoseEmbroider and the Aligner, here obtained from image input for training. We use the standard iterative residual network from [35] to predict the joint rotations from the mean parameters, and an MLP to regress shape coefficients.

In Table 3, we report the average pa-MPJPE (Procrustes-aligned Mean Per Joint Position Error) on BEDLAM-Script and 3DPW [68]. We note that the state-of-the-art SMPLer-X Huge model [10] achieves 43 and 41mm for each respectively, and that our PoseEmbroider with the trained SMPL-head is only subpar by 6 and 12mm, while (1) leveraging a smaller ViT (base) for encoding,

**Table 3: SMPL regression results for different representations and inputs.** The regression head is trained solely on BEDLAM-Script, with frozen image-based features of the Aligner/PoseEmbroider models. We report the pa-MPJPE in mm with the ground truth pose, on BEDLAM-Script (validation set) and 3DPW [68] (test set).

| pa-MPJPE↓                         | BEDLAM-Script | 3DPW |
|-----------------------------------|---------------|------|
| Aligner (image)                   | 50            | 54   |
| PoseEmbroider (image)             | 49            | 53   |
| <b>PoseEmbroider (image+text)</b> | <b>44</b>     | -    |



**Fig. 7: Image-based SMPL regression with an optional text hint.**

(2) not training the input representation end-to-end, and (3) training the regression head on 50K synthetic samples only (thus exposing the domain gap on 3DPW). Interestingly, our model improves by +11% when provided text cues. The PoseEmbroider design makes it possible to process added textual information without any retraining, *e.g.* to refine estimations as illustrated in Figure 7.

## 7 Discussion

**Conclusion.** We have introduced the PoseEmbroider framework, which derives visual-, 3D-, semantic-aware pose representations. Instead of *aligning* to fit *shared* information across modalities, it is trained to *combine* and thus *enrich* single-modality pose representations. Beyond its direct use in any-to-any multi-modal retrieval, the proposed versatile representation can be leveraged for complex downstream tasks requiring fine-grained human pose understanding, such as pose instruction generation or SMPL regression.

**Limitations and future work.** Future improvement could come from employing more aggressive losses (*e.g.* attempting to predict target features instead of solely learning to match them), training on more data (the 50k training samples pale in comparison to the 400 millions pairs CLIP was trained on), or incorporating a broader set of modalities (depth maps, 2D keypoints *etc.*). The described training procedure resorts to a single tri-modal dataset. Yet, we can envision learning from a set of uni-modal and bi-modal datasets, each coming with different groups of modalities.

**Acknowledgments.** This work is supported by the Spanish government with the project MoHuCo PID2020-120049RB-I00, and by NAVER LABS Europe under technology transfer contract ‘Text4Pose’.

## References

1. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 3DV (2019)
2. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In: NeurIPS (2021)
3. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. In: NeurIPS (2020)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
5. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop (2005)
6. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: CVPRW (2018)
7. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: CVPR (2023)
8. Bogu, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016)
9. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
10. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., et al.: Smpler-x: Scaling up expressive human pose and shape estimation. In: NeurIPS (2024)
11. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In: CVPR (2023)
12. Chen, Z., Li, Q., Wang, X., Yang, W.: Liftedcl: Lifting contrastive learning for human-centric perception. In: ICLR (2022)
13. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality (2023), <https://lmsys.org/blog/2023-03-30-vicuna/>
14. Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., Ouyang, W.: Unihcp: A unified model for human-centric perceptions. In: CVPR (2023)
15. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: PoseScript: 3D Human Poses from Natural Language. In: ECCV (2022)
16. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: PoseFix: Correcting 3D Human Poses with Natural Language. In: ICCV (2023)
17. Ding, Y., Tian, C., Ding, H., Liu, L.: The clip model is secretly an image-to-prompt converter. In: NeurIPS (2024)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
19. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K.,

- Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model. In: arXiv preprint arXiv:2303.03378 (2023)
20. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives (2018)
  21. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Chatpose: Chatting about 3d human pose. In: CVPR (2024)
  22. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: AIFit: Automatic 3D human-interpretable feedback models for fitness training. In: CVPR (2021)
  23. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NeurIPS (2013)
  24. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: ICCV (2021)
  25. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)
  26. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A Single Model for Many Visual Modalities. In: CVPR (2022)
  27. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa\*, A., Malik\*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)
  28. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
  29. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
  30. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall PTR (1994)
  31. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
  32. Hong, F., Pan, L., Cai, Z., Liu, Z.: Versatile multi-modal pre-training for human-centric perception. In: CVPR (2022)
  33. Ibrahim, S., Sun, X., Wang, P., Garg, A., Sanan, A., Omar, M.: Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In: ICCV (2023)
  34. Jin, Z., Hayat, M., Yang, Y., Guo, Y., Lei, Y.: Context-aware alignment and mutual masking for 3d-language pre-training. In: CVPR (2023)
  35. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
  36. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR (2022)
  37. Kim, H., Zala, A., Burri, G., Bansal, M.: FixMyPose: Pose correctional captioning and retrieval. In: AAAI (2021)
  38. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. In: AAAI (2023)
  39. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
  40. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
  41. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
  42. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
  43. Kwon, G., Cai, Z., Ravichandran, A., Bas, E., Bhotika, R., Soatto, S.: Masked vision and language modeling for multi-modal representation learning. In: ICLR (2023)

44. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)
45. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. In: NeurIPS (2023)
46. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: CVPR (2023)
47. Lin, Y., Wei, C., Wang, H., Yuille, A., Xie, C.: Smaug: Sparse masked autoencoder for efficient video-language pre-training. In: ICCV (2023)
48. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
49. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG (2015)
50. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: PoseGPT: Quantizing human motion for large scale generative modeling. In: ECCV (2022)
51. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: ICCV (2019)
52. Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4m: Massively multimodal masked modeling. In: NeurIPS (2024)
53. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: CVPR (2021)
54. Müller, M., Arzt, A., Balke, S., Dorfer, M., Widmer, G.: Cross-modal music retrieval and applications: An overview of key methodologies. IEEE Signal Processing Magazine (2018)
55. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
56. Papineni, K., Roukos, S., Ward, T., Jing Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
57. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)
58. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: ECCV (2022)
59. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: ICCV (2023)
60. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data (2016)
61. 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: ICML (2021)
62. Ruan, Y., Lee, H.H., Zhang, Y., Zhang, K., Chang, A.X.: Tricolo: Trimodal contrastive loss for text to shape retrieval. In: WACV (2024)
63. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
64. Tang, S., Chen, C., Xie, Q., Chen, M., Wang, Y., Ci, Y., Bai, L., Zhu, F., Yang, H., Yi, L., et al.: Humanbench: Towards general human-centric perception with projector assisted pretraining. In: CVPR (2023)
65. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: ECCV (2022)
66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
67. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: CVPR (2019)

68. Von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
69. Wang, Y., Wu, Y., Tang, S., He, W., Guo, X., Zhu, F., Bai, L., Zhao, R., Wu, J., He, T., et al.: Hulk: A universal knowledge translator for human-centric tasks. arXiv preprint arXiv:2312.01697 (2023)
70. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)
71. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)
72. Yin, K., Zou, S., Ge, Y., Tian, Z.: Tri-modal motion retrieval by learning a joint embedding space. arXiv preprint arXiv:403.00691 (2024)
73. Youwang, K., Ji-Yeon, K., Oh, T.H.: Clip-actor: Text-driven recommendation and stylization for animating human meshes. In: ECCV (2022)
74. Yuan, J., Zhang, X., Zhou, H., Wang, J., Qiu, Z., Shao, Z., Zhang, S., Long, S., Kuang, K., Yao, K., et al.: Hap: Structure-aware masked image modeling for human-centric perception. In: NeurIPS (2023)
75. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: ECCV (2022)
76. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)

## Supplementary Material

In this supplementary material, we present examples of tri-modal data samples from BEDLAM-Script and BEDLAM-Fix (Section A), and complete the explanations given in the main paper about these augmentations of BEDLAM (*e.g.* image selection criteria). We provide additional qualitative results and analysis (including limitations) of our models in Section B. The original annotations associated to the queries presented in the main paper are available in Section C. Finally, we give implementation details in Section D and discuss responsibility to human subjects (Section E).

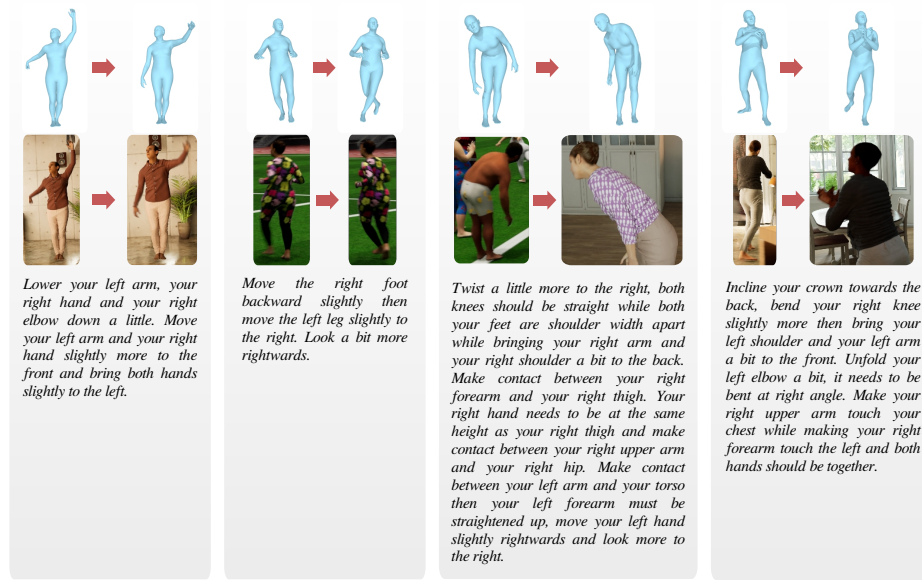
### A More about BEDLAM-Script and BEDLAM-Fix

**Dataset examples.** Figure A1 shows some examples of tri-modal samples from BEDLAM-Script. Figure A2 does the same for BEDLAM-Fix.



**Fig. A1: Examples from BEDLAM-Script.** A text describes the 3D pose represented in the image.

**Additional details about data selection.** To select samples from BEDLAM, we begin by a filtering of the images based on joint visibility, then proceed with the farthest pose sampling algorithm described in the main text of this paper. Here is the list of criteria used to select people images:



**Fig. A2: Examples from BEDLAM-Fix.** A textual instruction explains how to go from element A (left side of the arrow) to element B (right side of the arrow). Elements A and B can be either images and/or 3D poses.

- At least 16 of the main body joints had to be within the image boundaries (although they could be subject to occlusion).
- The person was required to be at the forefront (*i.e.*, positioned closest to the front compared to other individuals in the same image). However, we relaxed this condition slightly by also considering individuals positioned further in the background, provided that at least 70% of their bounding box did not overlap with the bounding box of someone positioned closer to the front.
- At least one side of the human bounding box (upscaled by a factor 1.1) had to be more than 224 pixels.

Paired elements (poses, images) in BEDLAM-Fix are part of BEDLAM-Script.

## B Additional qualitative results

### B.1 Text instruction generation

The automatic generation of pose instructions has several applications. It makes it possible to give correctional feedback to a trainee by comparing their pose to a trainer's. It also allows automatic narration to accompany sport training videos. In Figure A3, we provide qualitative results on real-world images depicting pilates moves. These pictures were obtained from YouTube videos of pilates classes.

We notice that the text instructions globally fit the different situations, especially for sitting and standing poses (four last columns of the first row). In particular, we find that the model is able to distinguish a set of fine-grained body changes, about arm position (1st row, columns 2,3 and 5), leg or arm bending (1st row, column 4; 2nd row,



columns 1 and 3), body twist (2nd row, columns 2, 4 and 5), head rotation (1st row, middle example) and so forth.

However, there are still small perception mistakes (*e.g.* *first image of the first row*: the right foot is not exactly on the floor; *middle image of the first row*: the elbows should not really be bent; *fourth image of the first row*: the right hand is rather on the right calf than the right thigh, but the confusion probably comes from the fact that the right wrist appears to be touching the right thigh). One recurrent behavior is the output of instructions like “move body part X to the right/left”, which do not always seem to really apply, from a human perspective. This likely stems from the fact that the bodies in Figure A3 are only visible from the side – thus preventing a fair depth estimate.

Globally, the study of several qualitative examples reveals that the model struggles with lying down poses. For instance, in the middle example of the second row, it believes the ground is the left side of the image (hence “the left thigh and the right forearm need to be *parallel to the floor*”). This problem comes directly from the pose representation, and can be explained by the low number of lying down poses in the training data. The frequency of such poses in the training batches could be increased to mitigate this issue.

Other typical failure cases include when the person is only partially visible (*e.g.* lower body truncated by the image boundaries). In these conditions, the text generation model would often output average instructions about the legs and forget about the main differences of the upper body. One way to alleviate this limitation could be to *also* train the model on truncated images input (instead of 3D poses only), and to consider instructions that only mention differences about the visible body parts. Two main aspects of our method makes this conceivable. First, the PoseEmbroider can similarly treat image and 3D pose input: it provides a modality-agnostic representation to the text decoder. Second, the model can be trained efficiently on synthetic data, using the automatic pipeline from [16], which can be modified so as to produce instructions involving a specific set of body joints (*i.e.*, those visible in the images).

## B.2 Any-to-any retrieval

We show a few more examples of any-to-any retrieval in Figure A4. We see that our model produces reasonable results for different types of query and target modalities.

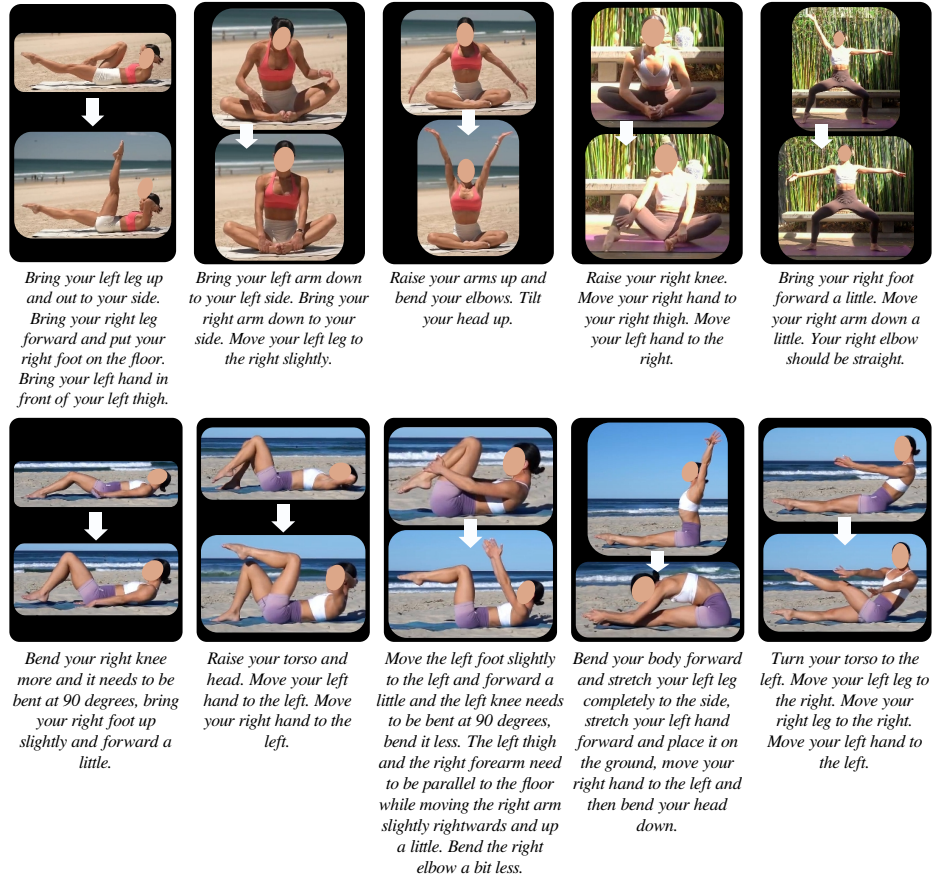
## C Original annotations for retrieval results

The original annotations for the queries presented in Figure 3 of the main paper were not displayed alongside the results; we provide them in Figure A5.

## D Implementation details

**Architecture details.** We detail below the architecture of each of our model components:

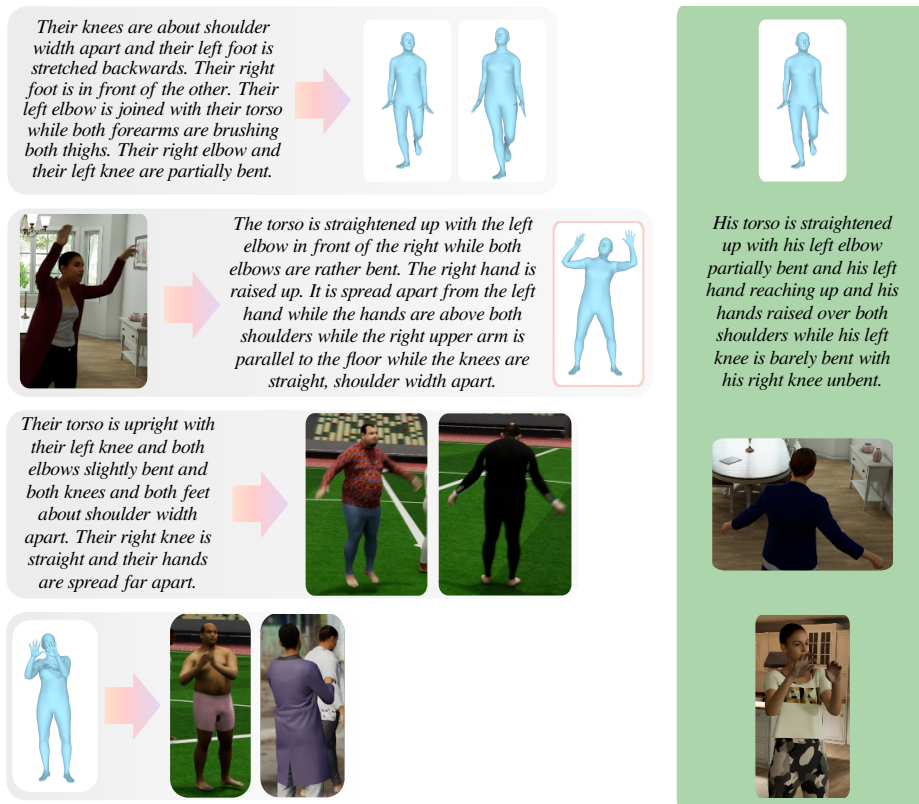
- *The pose encoder* is extracted from a Variational AutoEncoder (VAE) [40]. Its architecture follows VPoser [57], and was adapted to process the first 22 SMPL-X [57] body joint rotations (axis-angle representation). It shares the training objectives



**Fig. A3: Instruction generations on real-world images using the PoseEmbroider pose representation.** The instruction generation model was trained using the PoseEmbroider representations of 3D poses only. The generated text is shown below each image pairs. We occluded the faces to preserve privacy.

of the pose generative model in PoseFix [16], and has been trained on the 3D poses of BEDLAM-Script. For feature representation, we use the 512-dimensional vector that is further projected in the VAE to produce the distribution parameters. This frozen pretrained representation is fed to a trainable linear layer followed by a ReLU activation.

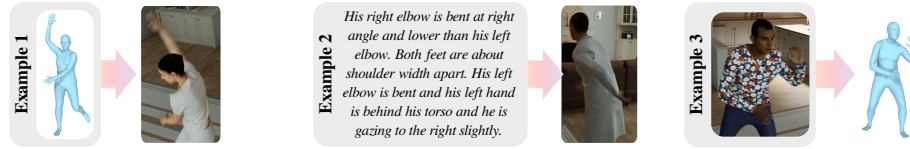
- *The text encoder* is the same as in [15]: text tokens are embedded thanks to a frozen DistilBERT [63], then fed to a transformer [66] (latent dimension 512, 4 heads, 4 layers, feed-forward networks of size 1024, GELU [31] activations, dropout rate of 0.1). The final single-vector embedding of a text is obtained by average-pooling all its token encodings. This frozen pretrained representation is further given to a trainable linear layer followed by a ReLU activation.
- *The image encoder* is the Vision Transformer [18] backbone from the SMPLer-X [57] base model, which was connected with a neural head and trained end-to-



**Fig. A4: Qualitative examples of any-to-any multi-modal retrieval** on the validation split of BEDLAM-Script. We show either top-1 or top-2 results for several types of single input and output modalities. Original paired modalities for the queries on the left are shown in the green box on the right. To ease reading, we additionally show the 3D pose associated retrieved texts and give it a pink border.

end for human mesh recovery. This image encoder is thus assumed to yield already powerful human-aware visual features and kept frozen. While SMPLer-X reasons on all image patches at once for SMPL regression, we average pool the visual tokens during the pretraining of the PoseEmbroider, based solely on synthetic data, and do not tune them in later stages (*e.g.* for SMPL regression, in Section 6 of the main paper). Specifically, the frozen, pretrained patch representations are aggregated into a single-vector representation after going through a trainable linear layer (projecting into a 512-dimensional space), and a ReLU activation.

- The *Embroider model* is a transformer [66] (latent dimension 256, 4 heads, 4 layers, feed-forward networks of size 512, GELU [31] activations, dropout rate of 0.1) followed by a LayerNorm [4]. It is sandwiched by two linear layers, projecting the input from a 512-dimensional space to the 256-dimensional working space of the transformer and vice-versa. Learned tokens (*i.e.*  $x$ ,  $e_v$ ,  $e_p$  and  $e_t$ ) are learnable parameters of size 512. Modality-specific reprojection MLPs, processing the PoseEmbroider output  $\bar{x}$ , consist in small multi-layer perceptrons [30] with 2 fully-



**Fig. A5: Original paired modalities** for the queries presented in Figure 3 of the main paper (qualitative any-to-any retrieval results).

connected layers of size 512 and a ReLU activation in-between. Their outputs are further L2-normalized.

- *The Aligner baseline model* appends modality-specific MLPs to each modality encoder. They consist of three linear layers with two in-between ReLU activations and dropout. Their hidden dimension is the same as the input, and they project into a 512-dimensional space.
- *The text decoder* of the text generation model has the same architecture as in PoseFix [16], except that it takes pose encodings of size 512 instead of 32. The pose encodings are fused with TIRG [67], projected thanks to a linear layer of dimension 512, then fed via cross-attentions to a transformer decoder (latent dimension 512, 8 heads, 4 layers, feed-forward networks of size 1024, GELU [31] activations, dropout rate of 0.1), which takes 512-dimensional token encodings as input. The output tokens are given to a linear layer of the size of the vocabulary to predict the likelihood of each subsequent word.
- *The SMPL regressor* relies on an iterative residual network as in [35]. We first concatenate the expected 512-sized input and the latest pose estimation (in the continuous 6D representation [76]), then feed the outcome to a 3-layer MLP (hidden dimension: 1024), evenly stuffed with dropout (default rate of 0.5) and Leaky ReLU activations. The results is further interpreted as the next pose estimate. The SMPL regressor also comprises a small 2-layer MLP (hidden dimension: 512; ReLU activation), taking the 512-dimensional PoseEmbroider feature as input to predict 10 shape coefficients.

The PoseEmbroider and Aligner models have a similar size of 164.8M parameters (even though the reprojection heads of the PoseEmbroider are expandable), including 162.5M just for the encoders.

**Optimization and training details.** The PoseEmbroider model is trained for 350 epochs, with all the uni-modal encoders frozen. We use mini-batches of size 128, a learning rate of  $2 \cdot 10^{-4}$ , the Adam [39] optimizer and a learning rate scheduler considering steps of size 400 and a gamma value of 0.5.

The text generation model is optimized with Adam, with a learning rate and weight decay of  $10^{-4}$ , for 900 epochs, and with batch sizes of 64. The finetuning on the PoseFix-OOS dataset and BEDLAM-Fix is run on 300 epochs. All trainings were done using precomputed cached features for the input pose representations.

The SMPL regression head undergoes a 100-epoch-long training with the same optimization hyper-parameters as for the text generation model.

**R-precision metrics for text generation.** This metric was originally proposed by [29] for motion-to-text generation, and directly imported by [16] for instruction text generation from pose pairs. It requires to first train an auxiliary retrieval model that links annotated texts and pose pairs. Then, for each generated text, this model is used to rank a set of pose pairs. The R-precision corresponds to the maximal rank of the

pose pair that was actually used to generate the text. [16] followed [29] and used a pool size of 32, however we report results on a harder pool size of 200.

**Text instruction generation model: comparison with PoseFix’s [16].** We explain here the differences with the setting in [16], which prevents the direct comparison of the models presented in this paper with those from [16]. First, [16] trains the pose encoder from scratch alongside the text decoder, which results in a pose encoder finetuned for the studied task. In this work, as we aim to compare off-the-shelf representations and offer inference from image input, we resort to pretrained frozen (and thus potentially sub-optimal) pose encoders, however allowing a mapping to the visual space. Next, while we both use data derived from AMASS [51] (recall that BEDLAM [7] uses motions from AMASS), it is not exactly the same. In particular, the construction of BEDLAM-Script and BEDLAM-Fix had to account for selection criteria on the images as well (joint visibility, inter-person occlusions, crop resolution *etc.*). In addition, in this work we only consider 50k poses (and 54k pairs) for pretraining, against 100k poses (and 95k pairs) in [16].

Aside from the aforementioned elements, the Aligner in Table 2 of the main paper is the closest to PoseFix’s original text generation model.

## E Responsibility to human subjects

Our models were trained exclusively on synthetic data from BEDLAM [7] and data from PoseFix [16] which includes human-written texts, but those do not carry any personal information.

The real-world images used to showcase the capabilities of our text generation model are solely used for qualitative studies. The Yoga images from Figure 6 of the main paper were obtained in studio with the written agreement of the subject. The images from Figure A3 were extracted from a public YouTube video, and we hide the faces to preserve anonymity.