PoseScript: 3D Human Poses from Natural Language – *Supplementary Material* –

Ginger Delmas[1,2], Philippe Weinzaepfel[2], Thomas Lucas[2], Francesc Moreno-Noguer[1], and Grégory Rogez[2]

[1] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2] NAVER LABS Europe

In Section 1, we present a description of the attached video; it is a live demo of our proposed text-conditioned generative model, trained on the human-written captions of PoseScript. We then provide additional information about the data collection process in Section 2. We give additional details on how to compute the different kinds of posecodes in Section 3, and specify a list of those that are used in our work. In Section 4, we elaborate on additional information about our automatic captioning pipeline and we compare different versions of the captions we produced. Additional statistics about our PoseScript dataset are presented in Section 5, and implementation details are given in Section 6.

1 Live demonstration of text-conditioned pose generation

The attached video shows a live demo of our text-conditioned generative model, pretrained on automatic captions and finetuned on human-written captions. The top part is a text area where the user can write a pose description, and generated poses that correspond to this query are displayed below. The demo starts with an empty text, and random sampled poses, *i.e.*, with latent variables randomly sampled from the standard normal distribution $\mathcal{N}_0$ (see Figure 1). The latency in showing the results is mainly due to the internet connection. It takes a few milliseconds to generate the generated poses, and a few more to visualize them. The video shows several generated poses per caption. Most of the constraints specified in the text description are satisfied by the sampled poses; the set of poses is also generally very diverse, which shows that our model is able to handle the ambiguities left in the caption. We also observe that more precise descriptions lead to less variability in the output samples.

Physical plausibility. We do not explicitly enforce plausibility constraints on the model outputs. As in VPoser [3] (same architecture), these must be learned by the model from the data. Empirically we did not notice any 'monster', except for some cases of inter-penetration, which also happen in AMASS. However, an improved prior could be trained with more data, *e.g.* with all poses from AMASS, with unconditional pose generation when no caption is available and text-to-pose generation when they are.

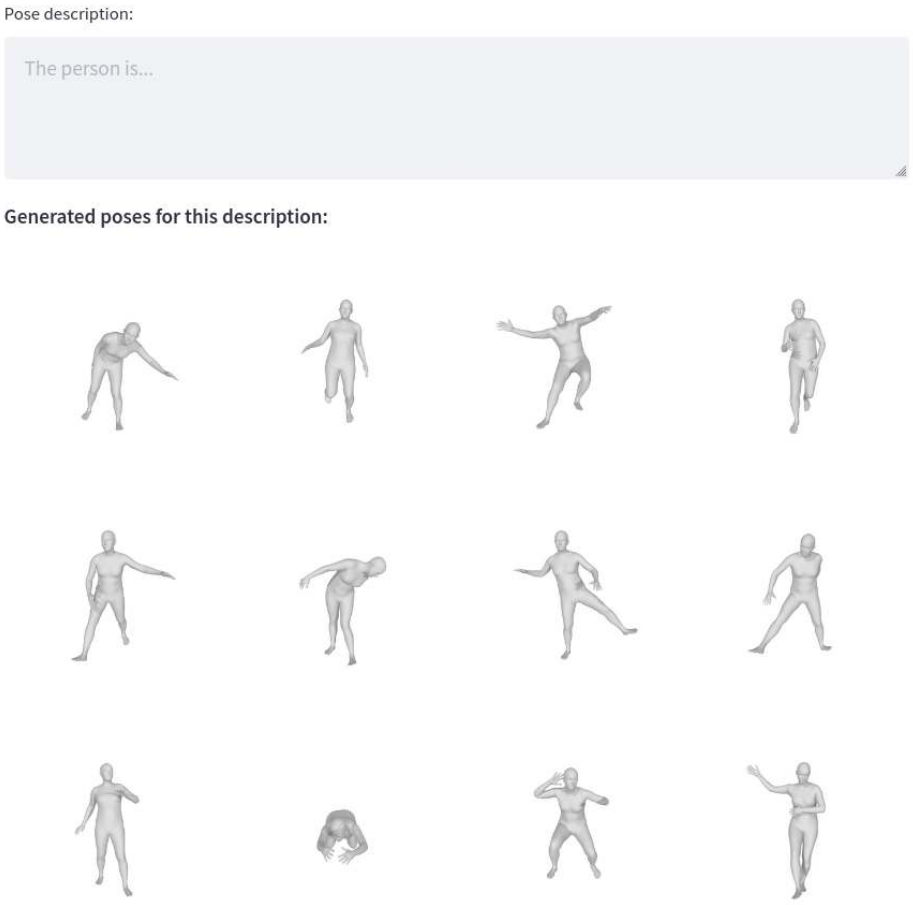
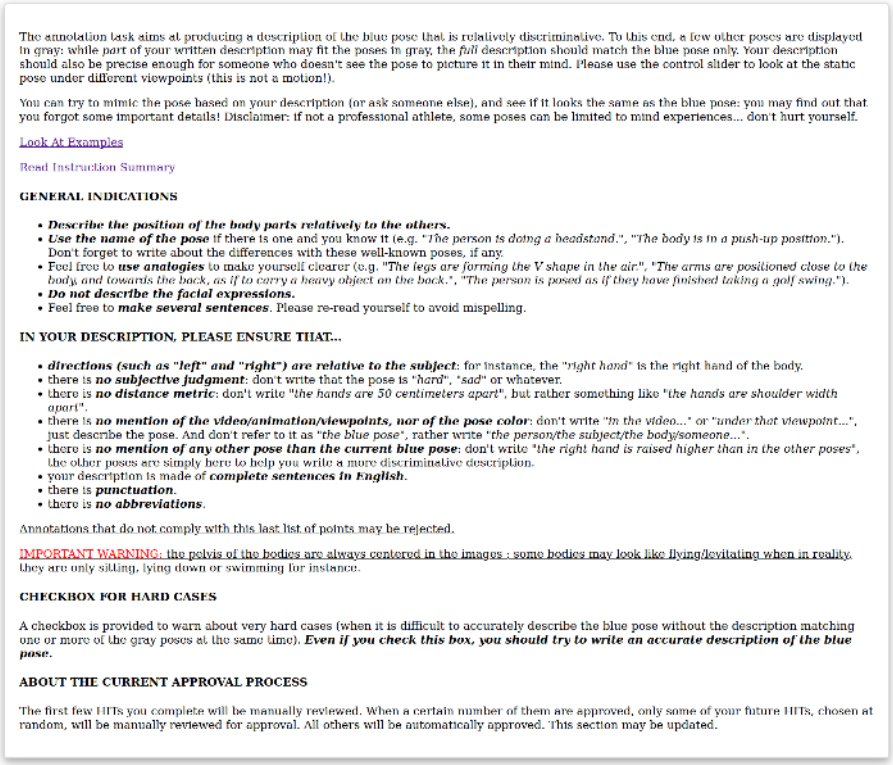


Fig. 1. Layout of the live demonstration of the text-conditioned pose generation model. See attached video.

2 Data collection process

Task instructions. A HIT (Human Intelligence Task) consists in writing the description of one given pose (in blue in the interface shown in Figure 3 of the main paper) precisely enough for the pose to be identified from its "discriminators" – the other similar poses, called discriminators (shown in grey in Figure 3 of the main paper). The instructions provided to the annotators are shown in Figure 2.



The annotation task aims at producing a description of the blue pose that is relatively discriminative. To this end, a few other poses are displayed in gray: while *part* of your written description may fit the poses in gray, the *full* description should match the blue pose only. Your description should also be precise enough for someone who doesn't see the pose to picture it in their mind. Please use the control slider to look at the static pose under different viewpoints (this is not a motion!).

You can try to mimic the pose based on your description (or ask someone else), and see if it looks the same as the blue pose: you may find out that you forgot some important details! Disclaimer: if not a professional athlete, some poses can be limited to mind experiences... don't hurt yourself.

Look At Examples

Read Instruction Summary

GENERAL INDICATIONS

- ***Describe the position of the body parts relatively to the others.***
- ***Use the name of the pose*** if there is one and you know it (e.g. *"The person is doing a headstand."*, *"The body is in a push-up position."*). Don't forget to write about the differences with these well-known poses, if any.
- Feel free to ***use analogies*** to make yourself clearer (e.g. *"The legs are forming the V shape in the air."*, *"The arms are positioned close to the body, and towards the back, as if to carry a heavy object on the back."*, *"The person is posed as if they have finished taking a golf swing."*).
- ***Do not describe the facial expressions.***
- Feel free to ***make several sentences***. Please re-read yourself to avoid mispelling.

IN YOUR DESCRIPTION, PLEASE ENSURE THAT...

- ***directions (such as "left" and "right") are relative to the subject***: for instance, the *"right hand"* is the right hand of the body.
- there is ***no subjective judgment***: don't write that the pose is *"hard"*, *"sad"* or whatever.
- there is ***no distance metric***: don't write *"the hands are 50 centimeters apart"*, but rather something like *"the hands are shoulder width apart"*.
- there is ***no mention of the video/animation/viewpoints, nor of the pose color***: don't write *"In the video..."* or *"under that viewpoint..."*, just describe the pose. And don't refer to it as *"the blue pose"*, rather write *"the person/the subject/the body/someone..."*.
- there is ***no mention of any other pose than the current blue pose***: don't write *"the right hand is raised higher than in the other poses"*, the other poses are simply here to help you write a more discriminative description.
- your description is made of ***complete sentences in English***.
- there is ***punctuation***.
- there is ***no abbreviations***.

Annotations that do not comply with this last list of points may be rejected.

IMPORTANT WARNING: the pelvis of the bodies are always centered in the images : some bodies may look like flying/levitating when in reality, they are only sitting, lying down or swimming for instance.

CHECKBOX FOR HARD CASES

A checkbox is provided to warn about very hard cases (when it is difficult to accurately describe the blue pose without the description matching one or more of the gray poses at the same time). ***Even if you check this box, you should try to write an accurate description of the blue pose.***

ABOUT THE CURRENT APPROVAL PROCESS

The first few HITs you complete will be manually reviewed. When a certain number of them are approved, only some of your future HITs, chosen at random, will be manually reviewed for approval. All others will be automatically approved. This section may be updated.

Fig. 2. Detailed task instructions provided to the annotators for the pose description task.

Selection of pose discriminators. To select the pose discriminators for a given pose to be annotated, we compare it to the other poses of PoseScript. Similarity is measured using the distance between their pose embeddings, with an early version of our retrieval model. Discriminators are required to be the closest poses, while having at least 15 different posecode categorizations. This ensures that the selected poses share some semantic similarities with the pose

to be annotated while having sufficient differences to be easily distinguished by the annotators. Discriminator examples are shown in Figure 3.

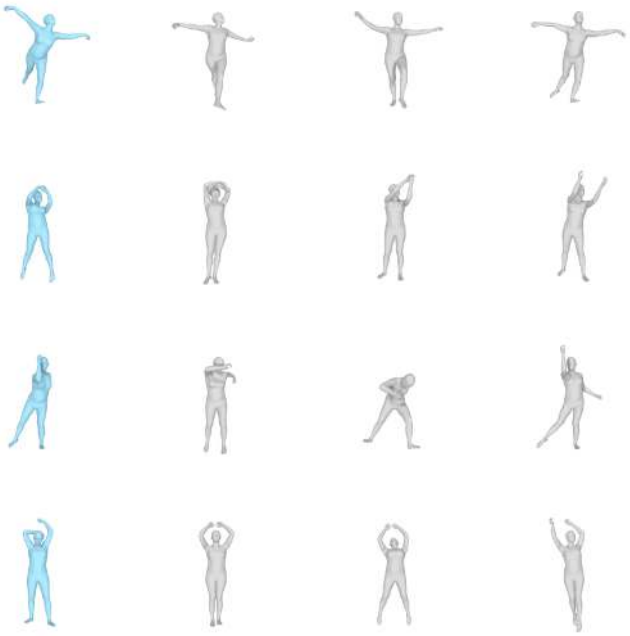


Fig. 3. Example of discriminators. For the pose shown in blue (left column) to be annotated, we show the three discriminators that were selected in grey.

Annotators qualifications. The HITs were initially made available for workers who:

- live in English-speaking countries (USA, Canada, Australia, United-Kingdom, New Zealand),
- got at least 5000 of their HITs approved in the past,
- already have an approval rate larger or equal to 95%.

We manually read and evaluated close to 1000 HITs, based on the following main criteria:

- The description is 'complete', *i.e.*, nearly all the body parts are described.
- There is no left/right confusion (early mistakes were tolerated and manually curated, as writing while assuming the point of view of the body pose is not an easy task).
- The description refers to a static pose, and not to a motion, as some people mistook the rotation of the bodies for a motion.

– There is no distance metric.
– There is no subjective comment regarding the pose.

Based on these criteria, we qualified workers who produced excellent descriptions, and in a second step, HITs were only made available to them.

The time to complete a HIT was estimated to be 2-3 minutes. Each HIT was rewarded $0.50, based on the minimum wage in California for 2022. We additionally paid $2 bonus to qualified annotators for every 50 annotations.

Statistics on the annotators. 159 different workers participated to the annotation process; 34 (21%) of them were qualified for the second annotation step. One annotator wrote 851 descriptions and five others close to or more than 300.

More human-written caption examples. To complement the human-written annotations shown in the main paper (left of Figure 2 of the main paper), we show in Figure 4 additional examples of human-written captions.

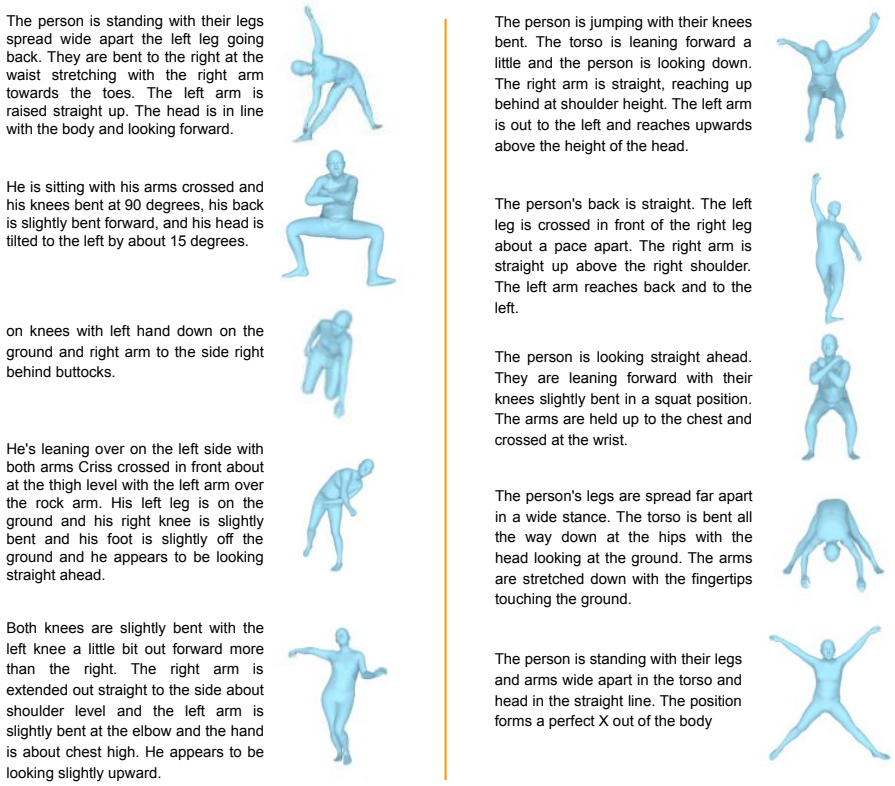


Fig. 4. Additional examples of human-written captions from the PoseScript dataset.

3 Posecodes

3.1 Computing posecodes

We detail here how the different kinds of posecodes are computed.

Elementary posecodes.

∘ *Angle posecodes* describe how a body part 'bends' around a joint j. Let a set of keypoints (i, j, k) where i and k are neighboring keypoints to j – for instance left shoulder, elbow and wrist respectively – and let p_l denote the position of keypoint l. The angle posecode is computed as the cosine similarity between vectors $v_{ji} = p_i - p_j$ and $v_{jk} = p_k - p_j$.

∘ *Distance posecodes* rate the $L2$-distance $\|v_{ij}\|$ between two keypoints i and j.

∘ *Posecodes on relative position* compute the difference between two sets of coordinates along a specific axis, to determine their relative positioning. A keypoint i is '`at the left of`' another keypoint j if $p_i^x > p_j^x$; it is '`above`' it if $p_i^y > p_j^y$; and '`in front of`' it if $p_i^z > p_j^z$.

∘ *Pitch & roll posecodes* assess the verticality or horizontality of a body part defined by two keypoints i and j. A body part is said to be '`vertical`' if the cosine similarity between $\frac{v_{ij}}{\|v_{ij}\|}$ and the unit vector along the y-axis is close to 0. A body part is said to be '`horizontal`' if it is close to 1.

∘ *Ground-contact posecodes* can be seen as specific cases of relative positioning posecodes along the y axis. They help determine whether a keypoint i is close to the ground by evaluating $p_i^y - \min_j p_j^y$. As not all poses are semantically in actual contact with the ground, we do not resort to these posecodes for systematic description, but solely for intermediate computations, to further infer super-posecodes for specific pose configurations.

Randomized binning step. As described above, each type of posecode is first associated to a value v (a cosine similarity angle or a distance), then binned into categories using predefined thresholds. In practice, hard deterministic thresholding is unrealistic as two different persons are unlikely to always have the same interpretation when the values are close to category thresholds, *e.g.* when making the distinction between '`spread`' and '`wide`'. Thus the categories are inherently ambiguous and to account for this human subjectivity, we randomize the binning step by defining a tolerable noise level η_τ on each threshold τ. We then categorize the posecode by comparing $v + \epsilon$ to τ, where ϵ is randomly sampled in the range $[-\eta_\tau, \eta_\tau]$. Hence, a given pose configuration does not always yield the exact same posecode categorization.

Super-posecodes are binary, and are not subject to the binning step. They only apply to a pose if all of the elementary posecodes they are based on possess the respective required posecode categorization.

3.2 List of posecodes

The list of the 77 elementary posecodes that are used in our work includes 4 angle posecodes, 22 distance posecodes, 34 posecodes describing relative posi-

tions (7 along the x-axis, 17 along the y-axis and 10 along the z-axis), 13 pitch & roll posecodes and 4 ground-contact posecodes. We specify the keypoints involved in the computation of each of these posecodes in Table 1. Conditions for posecode categorizations (*i.e.*, thresholds applied to the measured angles and distances, with the corresponding random noise level) are indicated for each kind of posecode in Table 2. Some of these elementary posecodes can be combined into super-posecodes. We list the 10 super-posecodes we currently consider in Table 3, and indicate for each of them the different ways they can be produced from elementary posecodes.

Angle posecodes
L-knee
R-knee
L-elbow
R-elbow

Ground-contact posecodes
L-knee
R-knee
L-foot
R-foot

Distance posecodes
L-elbow *vs.* R-elbow
L-hand *vs.* R-hand
L-knee *vs.* R-knee
L-foot *vs.* R-foot
L-hand *vs.* L-shoulder
L-hand *vs.* R-shoulder
R-hand *vs.* L-shoulder
R-hand *vs.* R-shoulder
L-hand *vs.* R-elbow
R-hand *vs.* L-elbow
L-hand *vs.* L-knee
L-hand *vs.* R-knee
R-hand *vs.* L-knee
R-hand *vs.* R-knee
L-hand *vs.* L-ankle
L-hand *vs.* R-ankle
R-hand *vs.* L-ankle
R-hand *vs.* R-ankle
L-hand *vs.* L-foot
L-hand *vs.* R-foot
R-hand *vs.* L-foot
R-hand *vs.* R-foot

Relative position posecodes
L-shoulder *vs.* R-shoulder (YZ)
L-elbow *vs.* R-elbow (YZ)
L-hand *vs.* R-hand (XYZ)
L-knee *vs.* R-knee (YZ)
R-foot *vs.* R-foot (XYZ)
neck *vs.* pelvis (XZ)
L-ankle *vs.* neck (Y)
R-ankle *vs.* neck (Y)
L-hip *vs.* L-knee (Y)
R-hip *vs.* R-knee (Y)
L-hand *vs.* L-shoulder (XY)
R-hand *vs.* R-shoulder (XY)
L-foot *vs.* L-hip (XY)
R-foot *vs.* R-hip (XY)
L-wrist *vs.* neck (Y)
R-wrist *vs.* neck (Y)
L-hand *vs.* L-hip (Y)
R-hand *vs.* R-hip (Y)
L-hand *vs.* torso (Z)
R-hand *vs.* torso (Z)
L-foot *vs.* torso (Z)
R-foot *vs.* torso (Z)

Pitch & roll posecodes
L-hip *vs.* L-knee
R-hip *vs.* R-knee
L-knee *vs.* L-ankle
R-knee *vs.* R-ankle
L-shoulder *vs.* L-elbow
R-shoulder *vs.* R-elbow
L-elbow *vs.* L-wrist
R-elbow *vs.* R-wrist
pelvis *vs.* L-shoulder
pelvis *vs.* R-shoulder
pelvis *vs.* neck
L-hand *vs.* R-hand
L-foot *vs.* R-foot

Table 1. List of elementary posecodes. We provide the keypoints involved in each of the posecodes, for each type of elementary posecodes (angle, distance, relative position, pitch & roll or ground-contact). We grouped posecodes on relative positions for better readability, as some keypoints are studied along several axes (considered axes are indicated in parenthesis). Letters 'L' and 'R' stand for 'left' and 'right' respectively. Ignored, skippable and unskippable posecodes are shown in Figures 5, 6, 7, 8, 9, 10 and 11.

Posecodes statistics. In Figures 5, 6, 7, 8, 9, 10 and 11 we show posecode statistics obtained over the 20,000 poses of the PoseScript dataset. Specifically, circle areas represent the proportion of poses satisfying the corresponding posecode categorization for the associated keypoints. We use the black and grey colors to denote categorizations that are ignored in the captioning process. A

Posecode type	Categorization	Condition
angle	`completely bent`	$v \pm 5 \leq 45$
	`almost completely bent`	$45 < v \pm 5 \leq 75$
	`bent at right angle`	$75 < v \pm 5 \leq 105$
	`partially bent`	$105 < v \pm 5 \leq 135$
	`slightly bent`	$135 < v \pm 5 \leq 160$
	`straight`	$v \pm 5 > 160$
distance	`close`	$v \pm 0.05 \leq 0.20$
	`shoulder width apart`	$0.20 < v \pm 0.05 \leq 0.40$
	`spread`	$0.40 < v \pm 0.05 \leq 0.80$
	`wide`	$v \pm 0.05 > 0.80$
relative position along the X axis	`at the right of`	$v \pm 0.05 \leq -0.15$
	`x-ignored`	$-0.15 < v \pm 0.05 \leq 0.15$
	`at the left of`	$v \pm 0.05 > -0.15$
relative position along the Y axis	`below`	$v \pm 0.05 \leq -0.15$
	`y-ignored`	$-0.15 < v \pm 0.05 \leq 0.15$
	`above`	$v \pm 0.05 > -0.15$
relative position along the Z axis	`behind`	$v \pm 0.05 \leq -0.15$
	`z-ignored`	$-0.15 < v \pm 0.05 \leq 0.15$
	`in front of`	$v \pm 0.05 > -0.15$
pitch & roll	`vertical`	$v \pm 5 \leq 10$
	`ignored`	$10 < v \pm 5 \leq 80$
	`horizontal`	$v \pm 5 > 80$
ground-contact	`on the ground`	$v \pm 0.05 \leq 0.10$
	`ground-ignored`	$v \pm 0.05 > 0.10$

Table 2. Conditions for posecode categorizations. The right column provides the condition for a posecode to have the categorization indicated in the middle column. v represents the estimated value (an angle converted in degrees, or a distance in meters), while the number after the $\pm$ denotes the maximum noise value that can be added to v. Thresholds and noise levels depend only on the type of posecode.

Subject	Configuration	Eligibility	Production
torso	`horizontal`	🟠	*pitch & roll* (pelvis, L-shoulder) = `horizontal` *pitch & roll* (pelvis, R-shoulder) = `horizontal`
body	`bent left`	🔵	*relativePos Y* (L-ankle, neck) = `below` *relativePos X* (neck, pelvis) = `at left` *or* *relativePos Y* (R-ankle, neck) = `below` *relativePos X* (neck, pelvis) = `at left`
body	`bent right`	🔵	*relativePos Y* (L-ankle, neck) = `below` *relativePos X* (neck, pelvis) = `at right` *or* *relativePos Y* (R-ankle, neck) = `below` *relativePos X* (neck, pelvis) = `at right`
body	`bent backward`	🟠	*relativePos Y* (L-ankle, neck) = `below` *relativePos Z* (neck, pelvis) = `behind` *or* *relativePos Y* (R-ankle, neck) = `below` *relativePos Z* (neck, pelvis) = `behind`
body	`bent forward`	🔵	*relativePos Y* (L-ankle, neck) = `below` *relativePos Z* (neck, pelvis) = `front` *or* *relativePos Y* (R-ankle, neck) = `below` *relativePos Z* (neck, pelvis) = `front`
body	`kneel on left`	🟠	*relativePos Y* (L-knee, R-knee) = `below` *ground-contact* (L-knee) = `on the ground` *ground-contact* (R-foot) = `on the ground`
body	`kneel on right`	🟠	*relativePos Y* (L-knee, R-knee) = `above` *ground-contact* (R-knee) = `on the ground` *ground-contact* (L-foot) = `on the ground`
body	`kneeling`	🟠	*relativePos Y* (L-hip, L-knee) = `above` *relativePos Y* (R-hip, R-knee) = `above` *ground-contact* (L-knee) = `on the ground` *ground-contact* (R-knee) = `on the ground` *or* *angle* (L-knee) = `completely bent` *angle* (R-knee) = `completely bent` *ground-contact* (L-knee) = `on the ground` *ground-contact* (R-knee) = `on the ground`
hands	`shoulder width apart`	🟠	*distance* (L-hand, R-hand) = `shoulder width` *pitch & roll* (L-hand, R-hand) = `horizontal`
feet	`shoulder width apart`	🔵	*distance* (L-foot, R-foot) = `shoulder width` *pitch & roll* (L-foot, R-foot) = `horizontal`

Table 3. List of super-posecodes. For each super-posecode, we indicate which body part(s) are subject to description (1st column) and their corresponding pose configuration (each super-posecode is given a unique category, indicated in the 2nd column). We additionally specify in the 3rd column whether the associated posecode is skippable for description, following the same color code as for elementary posecode statistics charts (🔵 : skippable; 🟠 : unskippable). Some super-posecodes can be produced by multiple sets of elementary posecodes: each set is separated by the word '*or*'. Letters 'L' and 'R' stand for 'left' and 'right' respectively.

black circle area means that the corresponding pose configuration is too ambiguous (*e.g.* when the relative distance between two body parts is close to 0, making the detection of the body parts' relative position less obvious.). Grey circle areas denote trivial pose configurations (*e.g.* when a left body part is at the left of the associated right body part: this is the case most of the time). They correspond to posecode categorizations that apply to at least 60% of the poses. In contrast, posecode categorizations that describe less than 6% of the poses are defined as unskippable (*i.e.*, such pose information cannot be randomly discarded during the posecode selection process), and are colored in orange. All other available posecodes categorizations, in blue, are skippable (*i.e.*, such pose information can be randomly discarded during the posecode selection process). Equivalent information for super-posecodes is provided in Table 3.

Most of the time, we follow statistics to consider posecode categorizations for pose description. In some specific cases, however, we are only interested in a subset of categorizations, and posecodes were only defined to retrieve such particular body pose information. This was done to infer super-posecodes later on (as for all ground-contact posecodes), or to bring in interesting semantics. For instance, distance posecodes involving one hand and another body part are only considered to inform about the position of the hand via the '`close`' category; indeed, while someone could describe the right hand as close to the left elbow, they are quite unlikely to point out that the right hand is wide apart from the left elbow. For the sake of completeness, we also present their statistics in the above-mentioned figures.

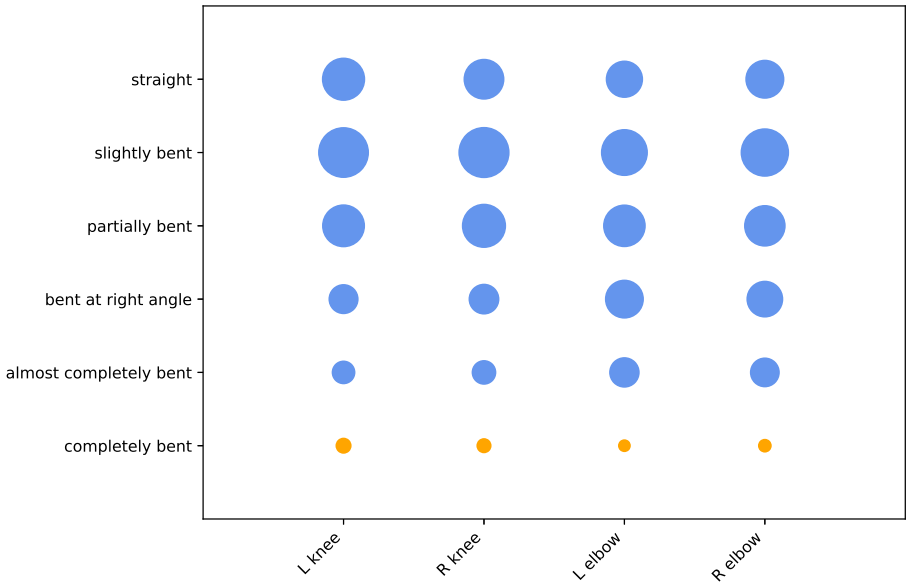


Fig. 5. Statistics on categorizations of angle posecodes, obtained over all the poses of the PoseScript dataset. Letters 'L' and 'R' refer to left and right body parts respectively. The dot size varies with the proportion of poses that fit to the given categorization. Posecode categorizations used at captioning time are represented in orange (unskippable) and blue (skippable). For any keypoint, the posecode interpretation '`completely bent`' applies to less than 6% of the poses and is hence defined as unskippable.

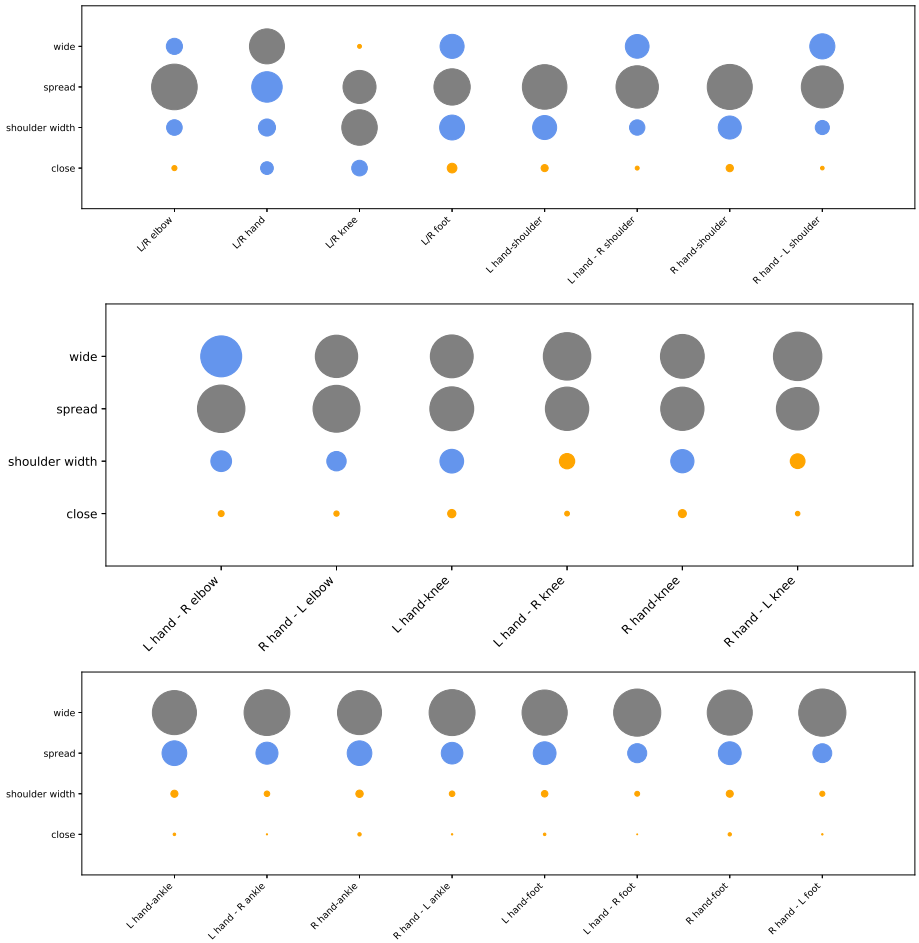


Fig. 6. Statistics on categorizations of distance posecodes, obtained over all the poses of the PoseScript dataset. The first four columns of dots from the top block show distance posecodes between the left and right corresponding body parts; other columns of dots study the distance between a left or right body part and another left or right body part (when the side of the second body part is not specified, it is the same as for the first body part). Letters 'L' and 'R' refer to left and right body parts respectively. The dot size varies with the proportion of poses that fit to the given categorization. The dot color indicates unskippable (orange), skippable (blue), and ignored (grey) posecodes, based on their scarcity. In practice, when a distance posecode involves one of the hands only, we just consider the '`close`' categorization.

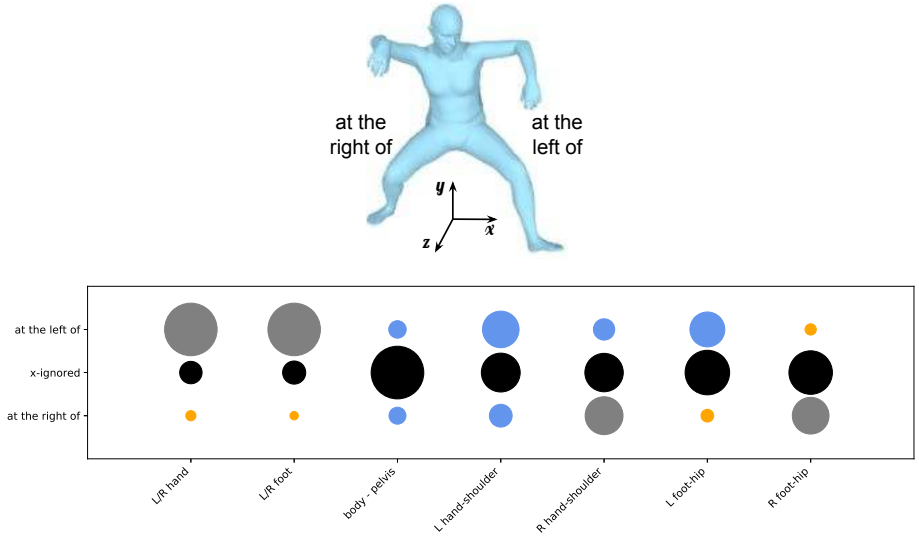


Fig. 7. Statistics on categorizations of relative position posecodes along the X axis, obtained over all the poses of the PoseScript dataset. Letters 'L' and 'R' refer to left and right body parts respectively. When unspecified, pairs of body parts are from the same side of the body. The dot size varies with the proportion of poses that fit to the given categorization. The dot color indicates unskippable (orange), skippable (blue), and ignored (grey) posecodes, based on their scarcity. Black dots are ignored because of their inherent ambiguity. For instance, it appears that, for less than 6% of the poses (orange dots), body extremities (hand, foot) are crisscrossed. Such posecode categorizations are rare, and hence defined as unskippable. In some rare cases, dots representing similar relations between left-only body parts and right-only body parts are of different colors (note that dot sizes are still similar) because numbers fall close to the thresholds defining whether a relation should be unskippable/skippable/ignored. In such cases, the same rule is applied for right and left relations, *i.e.*, the left hand (resp. foot) being at the left of the left shoulder (resp. hip) is considered to be a gray dot.

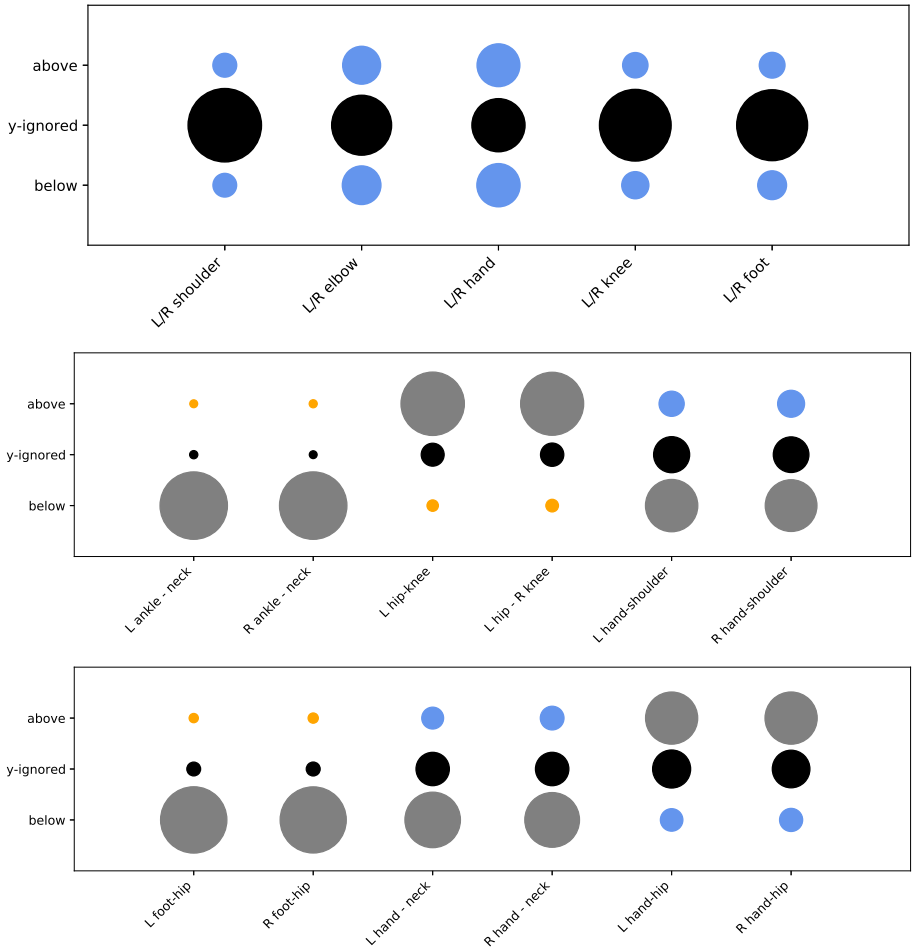


Fig. 8. Statistics on categorizations of relative position posecodes along the Y axis, obtained over all the poses of the PoseScript dataset. The top block shows the relative position of the left body part with respect to the corresponding right body part. Following blocks study other relations; when unspecified, pairs of body parts are from the same side of the body. Letters 'L' and 'R' refer to left and right body parts respectively. The dot size varies with the proportion of poses that fit to the given categorization. The dot color indicates unskippable (orange), skippable (blue), and ignored (grey) posecodes, based on their scarcity. Black dots are ignored because of their inherent ambiguity. Note that the dataset is quite balanced regarding left-related and right-related relations (similar dot sizes). Some of these posecodes are considered only for super-posecode inference (*e.g.* L ankle - neck); in such cases the scarcity matters less than the provided information.

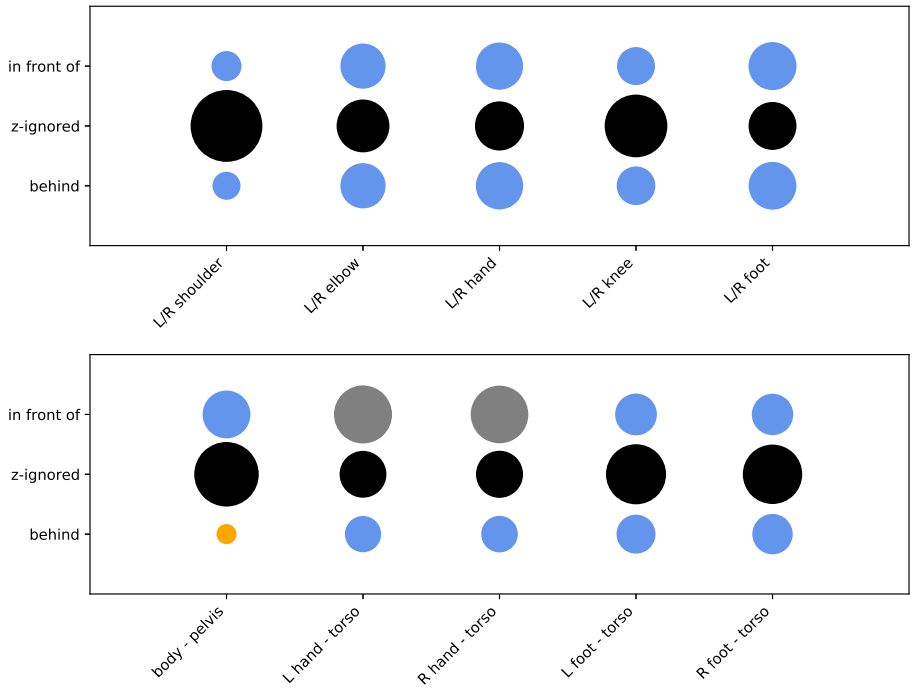


Fig. 9. Statistics on categorizations of relative position posecodes along the Z axis, obtained over all the poses of the PoseScript dataset. The top block shows the relative position of the left body part with respect to the corresponding right body part; the lower block mainly presents the relative position of body extremities (hand/foot) with respect to the torso. The first column of the lower block actually studies the position of the neck with regard to the pelvis to further determine whether the body is bent (forward/backward). Letters 'L' and 'R' refer to left and right body parts respectively. The dot size varies with the proportion of poses that fit to the given categorization. The dot color indicates unskippable (orange), skippable (blue), and ignored (grey) posecodes, based on their scarcity. Black dots are ignored because of their inherent ambiguity.

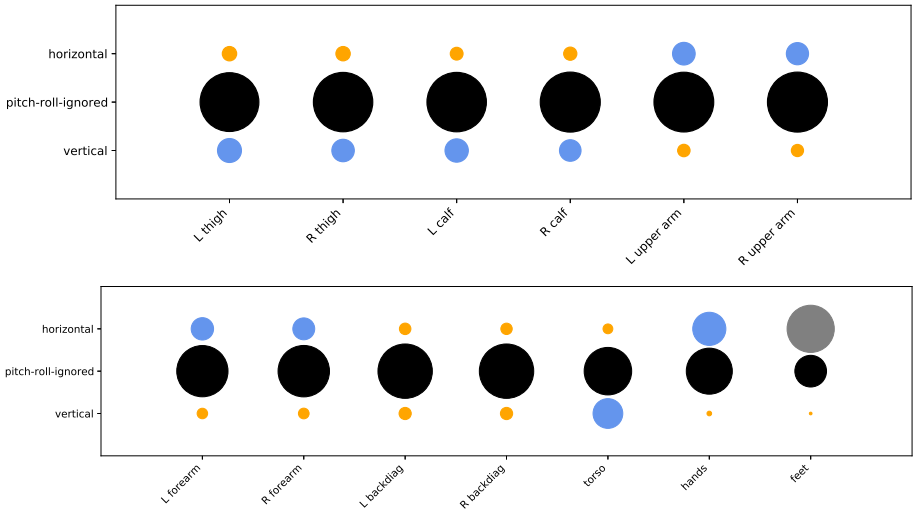


Fig. 10. Statistics on categorizations of pitch & roll posecodes, obtained over all the poses of the PoseScript dataset. Letters 'L' and 'R' refer to left and right body parts respectively. The word 'backdiag' refers to the segment between the pelvis and the shoulder, 'hands' (resp. 'feet') to the segment between the two hands (resp. feet), and 'torso' to the segment between the neck and the pelvis. The dot size varies with the proportion of poses that fit to the given categorization. The dot color indicates unskippable (orange), skippable (blue), and ignored (grey) posecodes, based on their scarcity. Black dots are ignored because of their inherent ambiguity. Some of these posecodes are considered only for super-posecode inference (*e.g.* hands horizontality); in such cases the scarcity matters less than the information provided.

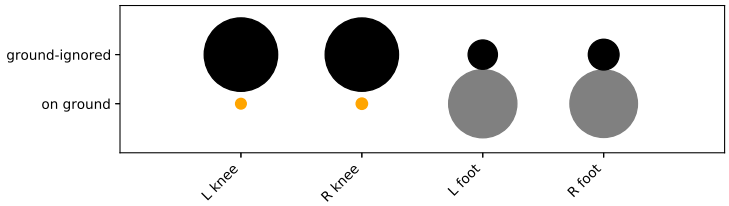


Fig. 11. Statistics on categorizations of ground-contact posecodes, obtained over all the poses of the PoseScript dataset. Letters 'L' and 'R' refer to left and right body parts respectively. The dot size varies with the proportion of poses that fit to the given categorization. While the dot colors indicate different levels of scarcity, the '`on the ground`' categorization is used for all of these posecodes independently, for super-posecode inference only.

4 More about the automatic captioning pipeline

In this section, we first detail the process used to generate the 6 automatic captions for each pose and report retrieval performance when pretraining on each of them and evaluating on human-written captions. Second, we present statistics about the captioning process. Third, we provide additional information about some steps of the captioning process.

4.1 Six versions of the automatically generated captions

All 6 captions, for each pose, were generated with the same pipeline. However, in order to propose captions with slightly different characteristics, we disabled some steps of the process when producing the different versions. Characteristics of the different caption versions are summarized in Table 4. Specifically, steps that were deactivated include:

- Removing redundant posecodes based on ripple effect rules.
- Adding a sentence constructed from high-level pose annotations given by BABEL [4].
- Implicitness, *i.e.*, aggregating posecodes; omitting support keypoints (*e.g.* 'the right foot is behind the torso' does not turn into 'the right foot is in the back' when this step is deactivated) ; randomly referring to a body part by a substitute word (*e.g.* 'it'/'they', 'the other').
- Randomly skipping eligible posecodes for description.

Among all 20k poses of PoseScript, only 6,628 are annotated in BABEL and may benefit from an additional sentence in their automatic description. As 39% of PoseScript poses come from DanceDB, which was not annotated in BABEL, we additionally assign the 'dancing' label to those DanceDB-originated poses, for one variant of the automatic captions that already leverages BABEL auxiliary annotations (see Table 4). This results in 14,435 poses benefiting from an auxiliary label. Figure 12 shows an example of each caption version for a given pose in PoseScript.

In Table 5, we report the retrieval performance on human-written captions when pretraining a retrieval model on each automatic caption version separately, then finetuning the models on human-written captions before evaluation. Results when pretraining on all the six caption versions together (*i.e.*, PoseScript-A) are provided in the last row (same as in Table 1 of the main paper). We first note that the best retrieval results were obtained in this setting, especially for the text-to-pose direction.

Next, we observe the impact of posecode aggregation and phrasing implicitness on retrieval performance by comparing results obtained by pretraining either on caption version `E` or on caption version `D`. Both caption versions share the same characteristics, except that version `E` is 'simplified'. This means that

Human-written caption

The person is like doing a pose of hip-hop dance. The body is leaning slightly to the left with the thighs close to the floor and with supports on the right heel, the left foot and the left hand. The right leg is forward with the knee slightly bent. The left leg is almost complemtely bent. The left arm is stretched vertically, a bit backward. The right arm is forward and slightly up.

Automatic caption [version A]

The person is in a dancing pose. The right hand is wide apart from the left hand, towards the sky, the left elbow is in the back of the right. The left shoulder is lower than the right shoulder, the thighs and the right upper arm are horizontal. The right elbow is in l-shape while the left knee is bent sharply, the right knee is partially bent, the right foot is front.

Automatic caption [version B]

A person is making a dance pose. The left knee is right next to the right knee. It is bent sharply. The right knee is partly bent, the right foot is in the front and in front of the left foot, the right arm is raised above the left with the right elbow in l-shape, the body is bent on the left side while the right hand is reaching up and wide apart from the left hand and the left arm is in the back of the other and the left shoulder is further down than the other, the left hand is in their back and both thighs and the right upper arm are horizontal.

Automatic caption [version C]

The person is bent on the left side while their left foot is behind the other. Their right upper arm and their thighs are aligned horizontally with their right knee bent and near their left knee while their right foot is to the front with their left knee completely bent. Their left arm is further down than their right arm and their hands are wide apart while their left elbow is straight and their right shoulder is further up than the left. Their left hand is reaching backward. It is beneath their left hip. Their right elbow is forming a l shape, their right arm is in front of the other.

Automatic caption [version D]

Their right shoulder is raised above the left and their right elbow is bent at near a 90 degree angle with their right arm wide apart from the other, their right hand is towards the sky. Their left arm is further down than their right arm. It is located behind their right arm. Their thighs are aligned horizontally and their left knee is completely bent and their right knee is partially bent with their left foot behind the right. This person is angled towards the left while their left hand is back, lower than their left hip with their right foot in the front.

Automatic caption [version E]

Their left hand is in the back of their torso with their right elbow forming a l shape with their right knee approximately shoulder width apart from their left knee with their right thigh parallel to the ground and their left thigh horizontal with their right knee bent while their right hand is raised higher than their right shoulder. Their left foot is located behind their right foot with their right foot located in front of their torso and the figure bent over with their right shoulder raised over their left shoulder. Their left elbow is further down than their right elbow while their left knee is bent sharply while their left hand is in the back of their right hand. Their left elbow is straight with their body inclined to the left side while their right upper arm is parallel to the ground.

Automatic caption [version F]

His right hand is spread far apart from his left hand with his left elbow unbent. His left elbow is underneath his right elbow with his left hand lower than his left hip with his left elbow located behind his right elbow with his right upper arm horizontal while his left thigh is parallel to the floor and his right shoulder is lying over his left shoulder while his right thigh is flat, his body is leaning forwards. His right elbow is at right angle with his right hand raised over his neck. His left hand is in the back of his torso while his left foot is located behind his right foot while his right foot is in front of his torso. His right hand is over his right shoulder, his right knee is rather bent. His right hand is raised higher than his left hand with the figure leaning on his left side with his left knee bent sharply. His left knee is at the same level as his right knee with his right hand ahead of his left hand.

Fig. 12. Captions from the different automatic versions for one pose in PoseScript.

Version	Random skip	Implicitness	Auxiliary labels	Ripple effect
A	✓	✓	✓(w/ dancing label)	✓
B	✓	✓	✓(w/ dancing label)	-
C	✓	✓	✓(w/o dancing label)	-
D	✓	✓	-	-
E	✓	-	-	-
F	-	-	-	-

Table 4. Summary of the automatic caption versions. ✓ symbols indicate when characteristics apply to each caption version.

E captions do not contain pronouns such as 'it' and 'the other', which represent an inherent challenge in NLP, as a model needs to understand to which entity these pronouns refer. Moreover, there is no omission of secondary keypoints (*e.g.* 'the right foot is behind the torso'). Hence, E captions have much less *phrasing* implicitness than D captions (note that there is still implicit *information* in the simplified captions, *e.g.* 'the right hand is close to the left hand' implicitly involves some rotation at the elbow or shoulder level). In Table 5, we observe a mean-recall increase of 2% when pre-training on E. This shows that aggregation and phrasing implicitness lead to more complex captions, as this is a source of error for cross-modal retrieval.

We then observe that using the additional 'dancing' label for poses originated from DanceDB does not bring any direct improvement (similar mean recall for B with respect to C), however it helps to reduce the variance: it may make it easier to distinguish between more casual poses (*e.g.* sitting) and highly various ones. Also, not using any BABEL label is better than using some, as evidenced by the 3.9 points difference between C and D). This can be explained by the fact that less than 33% of PoseScript poses are provided a BABEL label, and that those are too diverse (some examples include 'yawning', 'coughing', 'applauding', 'golfing'...) and too rare to robustly learn from. Importantly, many of these labels are motion labels and thus do not discriminate specific static poses.

Finally, we observe better performance when randomly skipping posecodes (E with regard to F), possibly because shorter and incomplete descriptions are closer to human-written captions (reduced domain gap). On the other hand, removing posecodes based on redundancy considerations does not seem to particularly help (A versus B).

4.2 Statistics about the captioning process

An average number of 303,495 'eligible' posecode categorizations were extracted from the 20,000 poses over the different caption versions (such 'eligible' posecodes are either represented by blue or orange dots in Figures 5, 6, 7, 8, 9, 10 and 11 for elementary posecodes, and in Table 3 for super-posecodes). During the posecode selection process, 42,981 of these were randomly skipped, and 6,286 were further removed to avoid redundancy. In practice, a bit less than 6% of the

	mRecall↑	pose-to-text			text-to-pose		
		R@1↑	*R*@5↑	*R*@10↑	*R*@1↑	*R*@5↑	*R*@10↑
test on the human captions (770 samples)							
pretrained on `A`	22.9±1.0	8.0±0.7	24.5±1.8	34.2±1.5	9.0±1.1	25.5±0.6	36.5±1.0
pretrained on `B`	23.9±0.5	9.1±0.4	25.2±1.4	36.2±1.4	9.6±0.3	25.7±0.4	37.4±0.3
pretrained on `C`	23.6±1.8	9.1±1.7	24.0±2.2	35.2±2.8	8.4±1.0	26.7±1.5	38.2±2.0
pretrained on `D`	27.5±2.3	10.2±1.9	29.1±3.1	40.0±2.5	11.1±1.8	31.5±2.6	43.4±2.7
pretrained on `E`	29.2±2.1	**12.0**±1.9	30.4±1.7	**42.8**±1.5	12.1±1.7	33.2±3.2	44.8±2.6
pretrained on `F`	26.8±0.6	10.5±0.8	27.5±0.5	39.8±0.6	10.6±0.8	30.2±1.4	42.6±0.9
pretrained on `A-F`	**30.4**±1.5	11.5±0.6	**32.1**±1.6	42.7±2.0	**12.6**±1.5	**35.4**±1.7	**48.0**±1.8

Table 5. Text-to-pose and pose-to-text retrieval results on the test split from PoseScript-H (human-written captions), when pretraining separately on each automatic caption version, then finetuning on PoseScript-H. The last row shows that pretraining on all the automatic captions together (`A-F`), *i.e.*, on PoseScript-A, yields the best performance, especially for text-to-pose retrieval. Results are reported as an average of 3 runs, with the standard deviations.

posecodes (17,570) are systematically kept for captioning due to being statistically discriminative (unskippable posecodes; orange dots). All caption versions were generated together in less than 6 minutes for the whole PoseScript dataset. Since the pose annotation task usually takes 2-3 minutes, it means we can generate 60k descriptions in the time it takes to manually write one.

Histograms about the number of posecodes used to generate the captions are presented in Figure 13. Automatic captions are based on an average number of 13.4 posecodes. Besides, we observed that less than 0.1% of the poses had the exact same set of 87 posecode categorizations than another.

Histograms about about the number of words per automatic caption are additionally shown in Figure 14.

4.3 Miscellaneous details

Input to the pipeline. The process takes 3D joint coordinates of human-centric poses as input. These are inferred using the neutral body shape with default shape coefficients and a normalized global orientation along the y-axis. We use the resulting pose vector of dimension $N \times 3$ ($N = 52$ joints for the SMPL-H model [5]), augmented with a few additional keypoints, such as the left/right hands and the torso. They are deduced by simple linear combination of the positions of other joints, and are included to ease retrieval of pose semantics (*e.g.* a hand is in the back if it is behind the torso). Specifically:

- the hand keypoint is computed as the center between the wrist keypoint and the keypoint corresponding to the second phalanx of the hand's middle finger.
- the torso keypoint is computed as the average of the pelvis, the neck and the third spine keypoint.

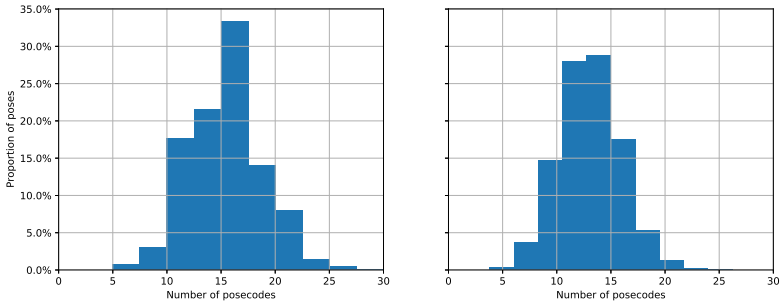


Fig. 13. Histograms of the number of posecodes used per caption. The left histogram presents the number of posecodes for the the caption version `F`, which does not perform random skipping. The number of posecodes of each pose, in the right histogram, was averaged over the other 5 caption versions produced for each given pose (the ripple effect rules were not yet applied for version `A`). Random skip reduces the number of posecodes and thus impacts the length of the caption (see Figure 14).

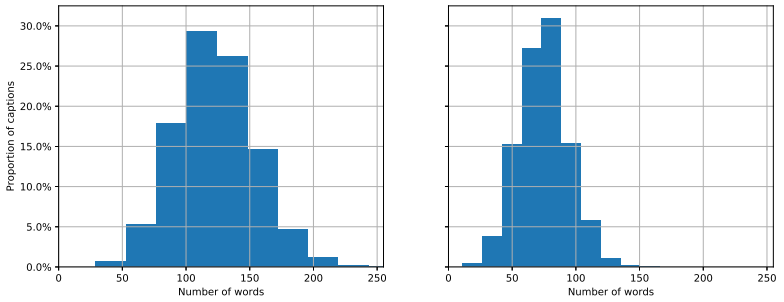


Fig. 14. Histograms showing the number of words per automatic caption, for version `F` (left) and version `D` (right). An explanation of the length difference is that version `D` was obtained by randomly skipping some posecodes and generally aggregating them. Version `D` captions are assumed to be closer to what humans would write.

What happens to posecodes contributing to super-posecodes?[3] There are three different outcomes for a posecode that contributes to a super-posecode:

- Some of the elementary posecodes are only *'support' posecodes*, and will never make it to the description alone: they only exist for computational purposes and need to be combined with other elementary posecodes to produce super-posecodes. For instance, to detect that the torso is parallel to the ground, we check that the two lines between the pelvis and each of the shoulders are horizontal. These two conditions are encoded via 'support' posecodes, which means that if the super-posecode is not produced because one of the two conditions is not satisfied, the second condition will not be transcribed in the caption: alone, it is meaningless.
- Some other posecodes can be considered as *'semi-support' posecodes*: they are discarded if the super-posecode they contribute to is successfully produced, but can make it to the description alone otherwise. For example, one way to detect that the body is kneeling is to check that both knees are completely bent, and in contact with the ground (otherwise the body could be in a squatting position). If all these conditions are met, the body is described as in a kneeling position and there is no need to further precise that the two knees are completely bent. If some of these conditions are not satisfied (*e.g.* the person is standing straight on their right foot), the super-posecode is not produced, and conversely to a 'support' posecode, the 'semi-support' posecode 'the left knee is completely bent' is not discarded, as it carries important information.
- Remaining elementary posecodes, which contribute to super-posecodes but are neither 'support' nor 'semi-support' posecodes will make it to the description, no matter whether the super-posecodes they contribute to can be produced or not – unless they are skipped down the road, of course.

For more information about which posecodes are support and semi-support posecodes, please refer directly to the code.

How is the redundancy tackled in the captions?[4] Posecodes are numerous, and yet encode a single body pose. Between these constraints and those intrinsic to the human body (*e.g.* arms attached to the torso by the shoulders), information overlap arises quickly. In the automatic captions, redundancy is tackled in several ways: (1) posecodes summarized in aggregation rules are removed: information is passed on, not duplicated; (2) most of the posecodes contributing to super-posecodes are 'support' posecodes, that exist only for super-posecode inference and are removed afterwards; (3) redundant posecodes are further removed thanks to two kinds of ripple effect rules: (i) rules based on statistically frequent pairs and triplets of posecodes, and (ii) rules based on transitive relations between body parts. In details:

[3] To reduce the verbosity of this paragraph, we refer to specific posecode categorizations as 'posecodes'.

[4] To reduce the verbosity of this paragraph, we refer to specific posecode categorizations as 'posecodes'.

– **Relation-based rules** are mined automatically for each pose, and applied before any aggregation rule. For a given pose, if we have 3 posecodes telling that $a < b$, $b < c$ and $a < c$ (with a, b, and c being arbitrary body parts, and $<$ representing a relation of order such as *'below'*), then we keep only the posecodes telling that $a < b$ and $b < c$, as it is enough to infer the global relation $a < b < c$. For instance, with both *'L hand in front of torso'* and *'R hand behind torso'*, the posecode *'L hand in front of R hand'* is removed.
– **Statistics-based rules.** Let X and Y be two sets of posecodes. Let's write $p \sim Z$ a pose p that has all posecodes in a given set Z. We define a statistics-based rule $X \Rightarrow Y$ (X 'implies' Y) if

$$\frac{\sum_{p \in \text{PoseScript}} p \sim (X \cup Y)}{\sum_{p \in \text{PoseScript}} p \sim X} \geq \tau, \tag{1}$$

with $\tau = 1$ (ideally). In other words, if all the poses which have posecodes $X \cup Y$ can be summarized as having X only, then any pose that has X necessarily would have Y. This is a relatively safe assumption, as poses from PoseScript were selected to be as diverse as possible. We automatically mined statistics-based rules $X \Rightarrow Y$ such that $size(X) \leq 2$ and $size(Y) = 1$ with the following considerations:

- the rule must involve eligible posecodes only, *i.e.*, posecodes that could make it to the description; trivial or ambiguous posecodes cannot be part of X or Y,
- the rule must be symmetrically eligible for the left and right sides: the rule must work the same for the whole body,
- the rule must affect at least 50 poses, *i.e.*, $\sum_{p \in \text{PoseScript}} p \sim X \geq 50$,
- the rule must hold for at least 80% of the PoseScript poses when $size(X) = 2$ (*i.e.*, $\tau = 0.8$) and 70% when $size(X) = 1$ ($\tau = 0.7$).

We further reviewed all mined rules manually, to keep only the most meaningful and dispose of the following:

- rules where one of the posecodes in X could be considered an 'auxiliary' posecode, *i.e.*, a posecode used only to select a smaller set and make the denominator in equation (1) small enough to get past the selection threshold τ. This is particularly obvious when Y and one of the X posecodes are about the upper body while the other X posecode is about the lower body, for instance.
- rules with weak conditions, *e.g.* when X posecodes are providing conditions on left body parts relatively to right parts, to derive in Y a 'global' result on left body parts.

Statistics-based rules are computed before but applied after entity-based and symmetry-based aggregation rules; they consist in removing the Y posecodes if they still exist. For instance, with *'L hand above shoulder'*, *'R hand below hip'*, the posecode *'L hand above R hand'* is removed.

As a side note, annotators were found to repeat themselves in some captions.

Entity-based aggregation. We defined two very simple entities: the arm (formed by the elbow, and either the hand or the wrist; or by the upper-arm and the

forearm) and the leg (formed by the knee, and either the foot or the ankle; or by the thigh and the calf).

Omitting support keypoints. We omit the second keypoint in the phrasing in those specific cases:

- a body part is compared to the torso,
- the hand is found 'above' the head,
- the hand (resp. foot) is compared to its associated shoulder (resp. hip), and is found either 'at the left of' or 'at the right of' of it. For instance, better than having 'the right hand is at the left of the left shoulder', which is quite tiresome, we would have *e.g.* 'the right hand is turned to the left'.

Use of negations in captions. We studied the use of negation in human-written captions: a bit less than 5% of them contain negations (*e.g.* '[close but] not touching' (20%), 'not quite/fully/completely/very' (15%), 'not bent' (10%)). Similar negations are easy to integrate in automatic caption templates. We did not include any as the proportion of negations in automatic captions would have been much greater than in human-written captions otherwise.

Context (environment/action) for pose generation. Context can be provided via another modality (*e.g.* an image) or freely expressed in natural language. We include BABEL [4] action labels in our automatic captions, and annotators were welcome to use analogies in their descriptions, *e.g. 'as if to climb a ladder'*. We primarily focus on learning explicit fine-grained relations between body parts (detailed & low-level pose indications). Physical environment constraints are beyond the scope of this work but make for an exciting future research direction.

Sensitivity to caption noise. We measure a variance of the mean recall below 0.5% when evaluating the retrieval model on 3 independent test sets obtained by generating different automatic captions per test pose, which shows robustness to changes in the query formulation. Some noise in human-written captions is inevitable but the generative model still produces reasonable results in practice.

5 Dataset statistics

In this section, we provide some additional statistics about the PoseScript dataset.

Pose selection. Poses were sampled from 14,096 AMASS [2] sequences. Specifically, the first and last 25 frames of each sequence were skipped to avoid initialization poses (*e.g.* T-poses). Then we sampled one pose every 25 to avoid getting too similar poses (*i.e.*, consecutive poses). We used farther-sampling to further select 20,000 poses, which were found to come from 3,306 different sequences. Figure 15 presents the AMASS sub-datasets from which come the poses selected for PoseScript. In particular, it appears that PoseScript poses come from almost all sequences of DanceDB and MPI_Limits that are available in AMASS; and that most of the poses in PoseScript actually come from DanceDB (39%), CMU (19%) and BioMotionLab (13%).

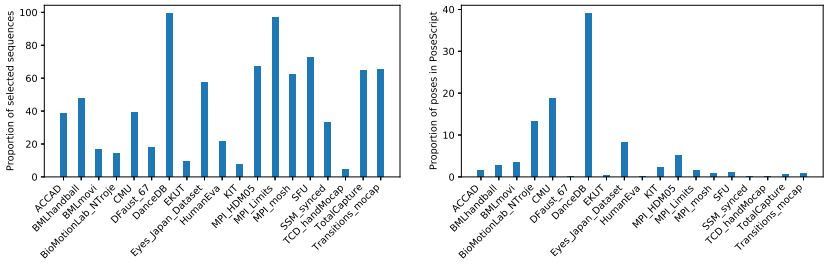


Fig. 15. Origin of the selected poses. The left bar plot shows the proportion of sequences that are eventually used in PoseScript with respect to available sequences in AMASS. A sequence is 'used' if it provided at least one pose to PoseScript. The right bar plot shows the distribution of the PoseScript poses over the AMASS sub-datasets.

Sequence-based split. The selected poses were split into 3 subsets, such that poses from the same sequence are allocated to the same subset. As a result, the train set contains 14,004 poses from 2,183 different sequences, the validation set contains 2,025 poses from 369 other different sequences, and the test set contains 3,971 poses from 754 other different sequences.

Human-written captions. Histograms about the number of tokens and the number of words per human-written caption are presented in Figure 16.

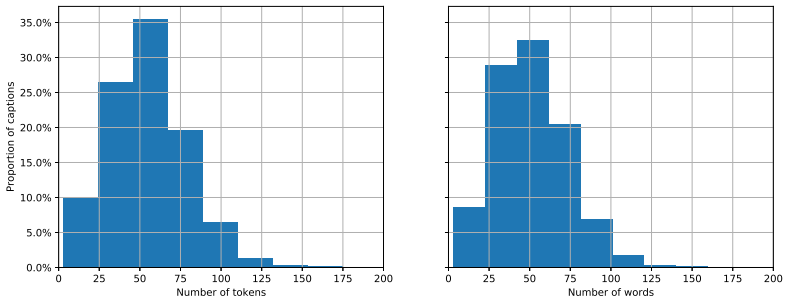


Fig. 16. Histograms showing the number of tokens (left) **and the number of words** (right) **per human-written captions**. Tokens include words plus punctuation.

6 Implementation details

Retrieval model. We use embeddings of size $d = 512$ and an initial loss temperature of $\gamma = 10$. GloVe word embeddings are 300-dimensional. The model is

trained end to end for 500 epochs, using Adam [1], a batch size of 32 and an initial learning rate of 2.10^{-4} with a decay of 0.5 every 20 epochs.

Generative model. We follow exactly VPoser [3] for the pose encoder and decoder architectures, and use the same text encoder as in the retrieval experiments. We train the models with a batch size of 128, using the Adam optimizer, a learning rate of 10^{-4} and a weight decay of 10^{-4}. The latent space has dimension 32.

References

1. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
2. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: ICCV (2019)
3. 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)
4. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: CVPR (2021)
5. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. In: SIGGRAPH Asia (2017)