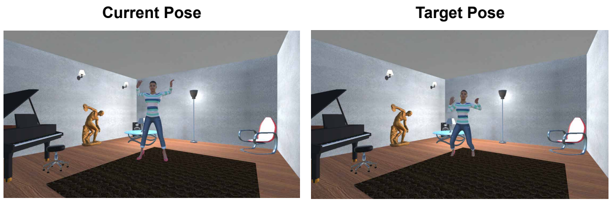
# FixMyPose: Pose Correctional Captioning and Retrieval

**Hyounghun Kim**[*], **Abhay Zala**[*], **Graham Burri**, **Mohit Bansal**
Department of Computer Science
University of North Carolina at Chapel Hill
{hyounghk, aszala, ghburri, mbansal}@cs.unc.edu

## Abstract

Interest in physical therapy and individual exercises such as yoga/dance has increased alongside the well-being trend, and people globally enjoy such exercises at home/office via video streaming platforms. However, such exercises are hard to follow without expert guidance. Even if experts can help, it is almost impossible to give personalized feedback to every trainee remotely. Thus, automated pose correction systems are required more than ever, and we introduce a new captioning dataset named FixMyPose to address this need. We collect natural language descriptions of correcting a "current" pose to look like a "target" pose. To support a multilingual setup, we collect descriptions in both English and Hindi. The collected descriptions have interesting linguistic properties such as egocentric relations to the environment objects, analogous references, etc., requiring an understanding of spatial relations and commonsense knowledge about postures. Further, to avoid ML biases, we maintain a balance across characters with diverse demographics, who perform a variety of movements in several interior environments (e.g., homes, offices). From our FixMyPose dataset, we introduce two tasks: the pose-correctional-captioning task and its reverse, the target-pose-retrieval task. During the correctional-captioning task, models must generate the descriptions of how to move from the current to the target pose image, whereas in the retrieval task, models should select the correct target pose given the initial pose and the correctional description. We present strong cross-attention baseline models (uni/multimodal, RL, multilingual) and also show that our baselines are competitive with other models when evaluated on other image-difference datasets. We also propose new task-specific metrics (object-match, body-part-match, direction-match) and conduct human evaluation for more reliable evaluation, and we demonstrate a large human-model performance gap suggesting room for promising future work. Finally, to verify the sim-to-real transfer of our FixMyPose dataset, we collect a set of real images and show promising performance on these images. Data and code are available: https://fixmypose-unc.github.io.



**Description 1 (English):** slide your right foot back one step and bend your knees, bring your wrists closer to your shoulders but maintain the position of your hands, finally drop your arms at the shoulder to level your hands with your neck.
**Description 2 (English):** bend both of your legs. bring both of your arms down almost below your ears. your left palm should be facing towards the chair. the back of your right hand should be facing the glass table.
**Description 3 (English):** bend both knees away from the lamp, lower down your body towards the rug, bring both hands down above your shoulder, right palm facing front and left palm facing the chair, tilt your head back a little towards the lamp.
**Description 1 (Hindi):** अपने दाहिने पैर को एक कदम पीछे खिसकाएं और अपने घुटनों को मोड़ें, अपनी कलाई को अपने कंधों के करीब लाएं लेकिन अपने हाथों की स्थिति को बनाए रखें, अंत में अपनी गर्दन के साथ अपने हाथों को समतल करने के लिए अपने हाथों को कंधे पर रखें।

Figure 1: Current and target image pair and the corresponding correctional descriptions in both English and Hindi (we show only one of the three Hindi descriptions due to space).

## 1 Introduction

As the well-being trend grows and people globally move to a new online lifestyle, interest in remotely (i.e., at home or in the office) learning health and exercise activities such as yoga, dance, and physical therapy is growing. Through advanced video streaming platforms, people can watch and follow the physical movements of experts, even without the expert being physically present (and hence scalable and less expensive). For such remote activities to be more effective, appropriate feedback systems are needed. For example, a feedback system should catch errors from the user's movements and give proper guidance to correct their poses. Related to this line of work, many efforts have been made on human pose estimation and action recognition (Johnson and Everingham 2010, 2011; Andriluka et al. 2014; Toshev and Szegedy 2014; Wei et al. 2016; Andriluka et al. 2018; Yan, Xiong, and Lin 2018; Zhao, Peng et al. 2019; Cao et al. 2019; Sun et al. 2019; Verma et al. 2020; Rong, Shiratori, and Joo 2020). Research on describing the difference between multiple images has also been recently active (Jhamtani and Berg-Kirkpatrick 2018; Tan et al. 2019; Park, Darrell, and Rohrbach 2019; Forbes et al. 2019). However, there has been less focus on the human pose-difference captioning tasks, which require solving unique challenges such as

[*] Equal contribution.

understanding spatial relationships between multiple body parts and their movements. Moreover, the reverse task of retrieving or generating a target pose is also less studied. Combining these two directions together can allow for more interweaving human-machine communication in future automated exercise programs.

Relatedly, interest in embodied systems for effective human-agent communication is increasing (Kim et al. 2018; Wang, Smith, and Ruiz 2019; Abbasi et al. 2019; Kim et al. 2020). Embodiment is also a desirable property when designing virtual assistants that provide feedback. For example, embodied virtual agents can show example movements to users or point at the users' body parts that need to move. Furthermore, for effective two-way communication with embodied agents, reverse information flow (i.e., human to agents) is also needed. A user may want to describe what actions they took so that the agent can confirm whether the user moved correctly or needs to change their movement. The agent should also be able to move its body to match the pose that the user is describing to help itself understand.

Therefore, to encourage the multimodal AI research community to explore these two tasks, we introduce a new dataset on detailed pose correctional descriptions called FIXMYPOSE (फिक्समाइपोज़), which consists of image pairs (a "current" and "target" image) and corresponding correctional descriptions in both English and Hindi (Fig. 1). To understand our dataset, imagine you are in a physical therapy program following an instructor in a prerecorded video at home. Your movements and resulting pose are likely to be wrong, hence, you would like a virtual AI assistant to provide detailed verbal guidance on how you can adjust to match the pose of the instructor. In this case, your incorrect pose is in the "current" image and the pose of the instructor is in the "target" image, forming a pair. The verbal guidance from the virtual AI assistant is the correctional description.

From our FIXMYPOSE dataset, we introduce two tasks for multimodal AI/NLP models: the 'pose-correctional-captioning' task and the 'target-pose-retrieval' task. In the pose-correctional-captioning task, models are given the "current" and "target" images and should generate a correctional description. The target-pose-retrieval task is the reverse of the pose-correctional-captioning task, where models should select the correct "target" image among other distractor images, given the "current" image and description. This two-task setup will test AI capabilities for both important directions in pose correction (i.e., agents generating verbal guidance for human pose correction, and reversely predicting/generating poses given instructions), to enable two-way communication between humans and embodied agents in future research. To generate image pairs, we implement realistic 3D interior environments (see Sec. 4 for details). We also extract body joint data from characters to allow diverse tasks such as pose-generation (Fig. 4). We collect descriptions for these image pairs by asking annotators from a crowdsourcing platform to explain to the characters how to adjust their pose shown in the "current" image to the one shown in the "target" image in an instructional manner from the characters' egocentric view (see Table 1). Furthermore, we ask them to refer to objects in the environment to create more detailed and accurate correctional descriptions, adding diversity and requiring models to understand the spatial relationships between body parts and environmental objects. The descriptions also often describe movement indirectly through implicit movement descriptions and analogous references (e.g., ”like you are holding a cane") (see Sec. 5.2), which means AI models performing this task should develop a commonsense understanding of these movements and references. To encourage multimodal AI systems to expand beyond English, we include Hindi descriptions as well (Fig. 1).

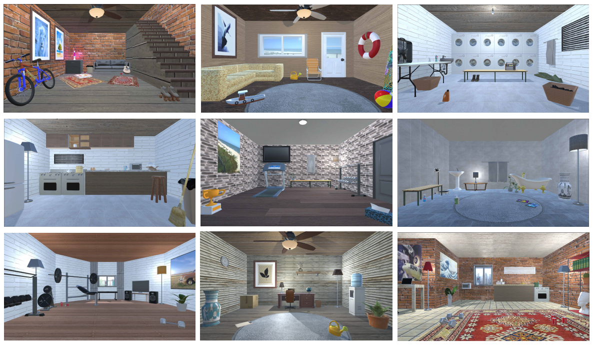


Figure 2: Example room environments: each room has a diverse style/theme (e.g., office, bathroom, living room).

Empirically, we present both unimodal and multimodal baseline models as strong starting points for each task, where we apply multiple cross-attention layers to integrate vision, body-joints, and language features. For the pose-correctional-captioning model, we employ reinforcement learning (RL), which uses self-critical sequence training (Rennie et al. 2017), for further improvement. Also, we present the results from a multilingual training setup (English+Hindi) which uses fewer parameters by sharing model components, but shows comparable scores.

The multimodal models in both tasks show better performance than unimodal models, across both qualitative human evaluation and several of the evaluation metrics, including our new task-specific metrics: object, body-part, and direction match (details in Sec. 8.1). There is also a large human-model performance gap on the tasks, allowing useful future work on our challenging dataset. We also show balanced scores on demographic ablations, implying that our dataset is not biased toward a specific subset. Furthermore, our model performs competitively with existing works when evaluated on other image-difference datasets (Image Editing Request (Tan et al. 2019), NLVR2 (Suhr et al. 2019), and CLEVR-Change (Park, Darrell, and Rohrbach 2019)). Finally, to verify the simulator-to-real transfer of our FIXMYPOSE dataset, we collect a test-real split which consists of real-world image pairs and corresponding descriptions, and show promising performance on the real images.

Our contributions are 3-fold: (1) We introduce a new dataset, FIXMYPOSE, to encourage research on the integrated field of human pose, correctional feedback systems on feature differences with spatial relation understanding, and embodied multimodal virtual agents; (2) We collect a multilingual (English/Hindi) dataset; (3) We propose two tasks based on our FixMyPose dataset (pose-correctional-captioning and target-pose-retrieval), and present several strong baselines as useful starting points for future work (and also demonstrate sim-to-real transfer).

## 2 Related Work

**Image Captioning.** Describing image contents in natural language has been actively studied (Xu et al. 2015; Yang et al. 2016; Rennie et al. 2017; Lu et al. 2017; Anderson et al. 2018a; Melas-Kyriazi, Rush, and Han 2018; Yao et al. 2018). This progress has been encouraged by the introduction of large-scale captioning datasets (Hodosh, Young, and Hockenmaier 2013; Lin et al. 2014; Plummer et al. 2015; Krishna et al. 2017; Johnson, Karpathy, and Fei-Fei 2016; Krause et al. 2017). Recently, more diverse image captioning tasks, which consider two images and describes the difference between them, have been introduced (Jhamtani and Berg-Kirkpatrick 2018; Tan et al. 2019; Park, Darrell, and Rohrbach 2019; Forbes et al. 2019). However, to the best of our knowledge, there exists no captioning dataset about describing human pose differences. Describing pose difference or body movement requires detailed multi-focus over all body parts and understanding relations between them, introducing new challenges for AI agents. This kind of dataset is promising because of its potential real-world applications in activities such as yoga, dance, and physical therapy.

**Human Pose.** Human pose estimation and action recognition have been a long-standing topic in the research community (Johnson and Everingham 2010, 2011; Andriluka et al. 2014; Toshev and Szegedy 2014; Wei et al. 2016; Andriluka et al. 2018; Yan, Xiong, and Lin 2018; Zhao, Peng et al. 2019; Cao et al. 2019; Sun et al. 2019; Verma et al. 2020; Rong, Shiratori, and Joo 2020). Recently, researchers are also focusing on generation tasks which generate a body pose sequence from an input of a different type from another modality such as audio or spoken language (Shlizerman et al. 2018; Tang, Jia, and Mao 2018; Lee et al. 2019; Zhuang et al. 2020; Saunders, Camgoz, and Bowden 2020). However, there have been no research attempts on text generation based on pose correction. Thus, our novel FixMyPose dataset will encourage the community to explore this new direction.

**Spatial Relationships.** Understanding spatial relationships between objects is an important capability for AI agents. Thus, the topic has attracted much attention from researchers (Bisk, Marcu, and Wong 2016; Wang, Liang, and Manning 2016; Li et al. 2016; Bisk et al. 2018). Our FixMyPose dataset is rich in such reasoning about spatial relations with a variety of expressions (not only simple directions of left/right/up/down). Moreover, all the spatial relationships in the descriptions of the FixMyPose dataset are considered from the characters' egocentric perspective, requiring models to understand the characters' viewpoints.

**Virtual Assistants.** Virtual AI assistants such as Alexa, Google Assistant, Cortana, and Siri are already ubiquitous in our lives. However, there has been an increasing demand for multimodal (i.e., vision+language) virtual AI assistants, and as robotic and virtual/augmented/mixed reality technologies grow, so does interest in embodied virtual assistants (Kim et al. 2018; Wang, Smith, and Ruiz 2019; Abbasi et al. 2019; Kim et al. 2020). Our FixMyPose dataset will contribute to the evolution of embodied multimodal virtual assistants by providing a novel dataset as well as proposing a new approach on how to integrate physical movement guidance with virtual AI assistants.

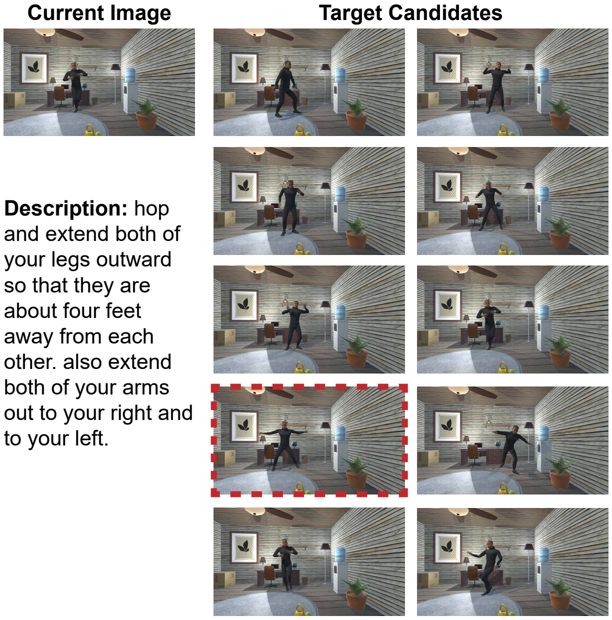

Current Image
Target Candidates
Description: hop and extend both of your legs outward so that they are about four feet away from each other. also extend both of your arms out to your right and to your left.

Figure 3: The target-pose-retrieval task: models have to select the correct "target" image from a set of distractors (the image with red dashed border is the ground-truth target pose), given "current" image and correctional description.

## 3 Tasks

**Pose Correctional-Captioning Task.** During this task, the goal is to generate natural language (NL) correctional descriptions, considering the characters' egocentric view, that describe to a character how they should adjust their pose shown in the "current" image to match the pose shown in the "target" image (Fig. 1). As the "current" and "target" image pairs contain various objects in realistic room environments, models should have the ability to understand the spatial relationships between the body parts of characters and the environment from the characters' perspectives.

**Target Pose Retrieval Task.** Here, the goal is to select the correct "target" image among 9 incorrect distractors, given the "current" image and the corresponding correctional description (Fig. 3). For the distractor images, we only consider images that are close to the "target" pose in terms of body joints distances (see Appendix in arxiv full version for detailed criteria). These distractor choices discourage models from easily discerning the correct "target" image via shallow inference or shortcuts, requiring minute differences to be captured by models. The large human-model perfor-

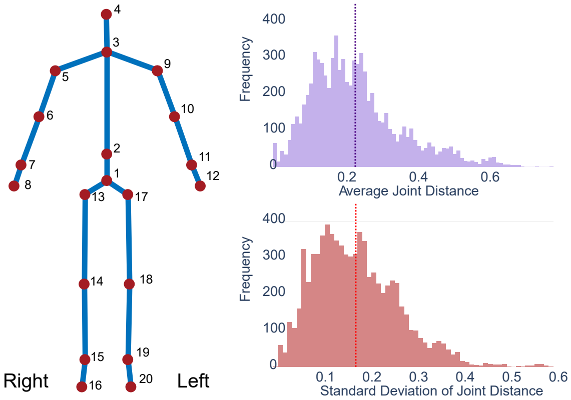


Figure 4: The 3D joint configuration of characters (left). The distribution of joint distances (meters) between poses of the "current" and "target" images (right). The Avg. of Min joint distances: 0.04 and the Avg. of Max joint distances: 0.65.

mance gap (Sec. 8.2) verifies the quality of our distractors.

## 4 Dataset

Our FixMyPose dataset is composed of image pairs with corresponding correctional descriptions in English/Hindi.

**Image, 3D Body Joints, and Environment Generation.** We create 25 realistic 3D diverse room environments, filled with varying items (Fig. 2). To ensure diversity, we employ 6 human character avatars of different demographics across gender/race (each character is equally balanced in our dataset).[1] Since creating/modifying the body of characters requires 3D modeling/artistic expertise, we use pre-made character models that are publicly available (hence also copyright-free for our community's future use) in Adobe's Mixamo[2]. In the rooms, the characters perform 20 movement animations and the camera captures images on a fixed interval. We also obtain 3D positional body joint data of the character's poses in the "current" and "target" images to provide additional useful features and allow a potential reverse pose-generation task (Fig. 4). See Appendix in arxiv full version for more on animation examples, environment creation, body joint data, and image capturing.

**Description Collection.** We employ annotators from the crowdsourcing platform Amazon Mechanical Turk[3] to collect the correctional descriptions. Workers are provided 3 images, "current", "target" images, and a "difference" image that shows the difference between the two images, allowing them to write clear descriptions (see Appendix in arxiv full version for the images and collection interface). We ask them to write as if they are speaking to the characters as assistants who are helping them (like "You should ..."), not calling them by the 3rd person (like "The person ...", "They/She/He ..."). It also helps prevent accidental biased terms assuming the demographics of the characters. We collect 1 description for each image pair for the train split and 3 for all subsequent splits (i.e., val-seen/val-unseen/test-unseen) from unique workers, making the computation of automated evaluation metrics such as BLEU possible.

| Reference Frame | Freq. | Example (English) |
|---|---|---|
| Egocentric Relation | 100% | "... **rotate your left shoulder** so that your hand is **above your elbow** ..." |
| Environmental Direction | 52% | "... turn your left leg and right leg to the left to **face the wall with the door** ..." |
| Implicit Movement Description | 58% | "... lean your body towards and slightly over your right leg ..." |
| Analogous Reference | 18% | "... in front of you **as if you are gesturing for someone to stop** ..." |

Table 1: Examples of linguistic properties in correctional descriptions (see Appendix in arxiv full version for examples and image examples of implicit movement description).

**Description Verification.** Each description and its corresponding image pair is given to a separate group of workers through a verification task. For each description, 3 different workers are asked to rank it from 1-4 based on its relevance to the image pair and its clarity, similar to previous works (Lei et al. 2020). Descriptions that 2/3 of the workers rate lower than 3 are discarded. Image pairs that are flagged with certain issues are discarded as they do not provide good data (see Appendix in arxiv full version for the interface).

**Hindi Data Collection.** To collect the translated Hindi descriptions, we present a translation task to workers. Workers are given a description that has passed the verification task and its corresponding image pair to ensure the original meaning is not lost (see Appendix in arxiv full version for the translation interface).

**Worker Qualification and Payment.** We require workers completing either of the tasks to be fluent in the needed languages and to have basic MTurk qualifications. The writing task takes around 1 minute and workers are paid $0.18 per description. To encourage workers to write more and better descriptions, an additional increasing-bonus system is implemented. See Appendix in arxiv full version for qualification/bonus/payment details.

## 5 Data Analysis

We collect 7,691 image pairs and 11,127 correctional descriptions for both English and Hindi (1 per train and 3 per evaluation splits). Our dataset size is comparable to other captioning tasks/datasets such as Image Editing Request (Tan et al. 2019) (3.9K image pairs/5.7K instructions), Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) (13.2K image pairs/captions), and Birds-to-Words (Forbes et al. 2019) (3.3K image pairs/16K paragraphs). We plan to keep extending the dataset and add other languages in the future.

[1] Our task focuses on understanding body movements/angles and not demographics, but we still ensure demographic diversity and balance in our dataset for ethical/fairness purposes so as to avoid unintended biases (e.g., see the balanced demographics ablation results and Sim-to-Real Transfer results on people with different demographics with respect to the 6 character avatars in Sec. 8). We plan to further expand our dataset with other types of diversity (e.g., height, age) based on digital avatar availability.

[2] https://www.mixamo.com

[3] https://www.mturk.com

### 5.1 Statistics

**Joint Distances.** Fig. 4 shows the distribution of average joint distances (meters) between the poses in the "current" and "target" images. As indicated by the mean (0.24m), std-dev (0.18m), and min/max (0.04/0.65m) of the average distance of individual joints, models should be able to capture different movement levels simultaneously in an image pair.

**Description Vocabulary and Length.** The collection of descriptions in our FIXMYPOSE dataset has 4,045/4,674 unique English/Hindi words. The most common words in both languages (see Appendix in arxiv full version for details and pie charts) relate to direction, body parts, and movement, showing that models need to have a sense of direction with respect to body parts and objects, and also capture the differences between the poses to infer the proper movements. The average length of the multi-sentenced descriptions (49.25/52.74 words) is high, indicating that they are well detailed (see Appendix in arxiv full version for details).

### 5.2 Linguistic Properties

To investigate the diverse linguistic properties in our dataset, we randomly sample 50 descriptions and manually count occurrences of traits. We found interesting traits (see Table 1 and Appendix in arxiv full version for examples), requiring agents to deeply understand characters' movements and express them in an applicable form (the Hindi descriptions also share these traits).

**Egocentric and Environmental Direction.** Descriptions in our FIXMYPOSE dataset are written considering the egocentric (first-person) view of the character. Descriptions also reference many environmental objects and their relation to the characters' body parts, again from an egocentric view. This means models must understand spatial relations of body parts and environmental features from the egocentric view of the character rather than the view of the "camera".

**Implicit Movement Description and Analogous Reference.** Implicit movement description and analogous reference are often present in descriptions. These descriptions imply movements without needing to say them. Analogous references are a more extreme form of implicit movement description, where the movement is wrapped in an analogy. Models must develop commonsense knowledge of these movements in order to understand their meaning. See Table 1 and Appendix in arxiv full version for examples.

## 6 Models

We present multiple strong baselines for both the pose-correctional-captioning and target-pose-retrieval task (Fig. 5) to serve as starting points for future work.

### 6.1 Pose Correctional Captioning Model

We employ an encoder-decoder model for the pose-correctional-captioning task. Also, we apply reinforcement learning (RL) after training the encoder-decoder model, and present multilingual training setup which reduces the number of parameters through parameter sharing.

**Encoder.** We employ ResNet (He et al. 2016) to obtain visual features from images. To be specific, we extract feature maps $f^c$ and $f^t \in \mathbb{R}^{N \times N \times 2048}$ from the "current" pose image $I^c$ and the "target" pose image $I^t$, respectively: $f^c = \text{ResNet}(I^c)$; $f^t = \text{ResNet}(I^t)$. For 3D joints, $J^c, J^t \in \mathbb{R}^{20 \times 3}$, we use linear layer to encode: $\hat{J}^c = \text{PE}(W_j^\top J^c)$; $\hat{J}^t = \text{PE}(W_j^\top J^t)$; $J^d = \text{PE}(W_j^\top (J^t - J^c))$, where $W_j$ is the trainable parameter (all $W_*$ from this point on denote trainable parameters) and PE (Gehring et al. 2017; Vaswani et al. 2017) denotes positional encoding.

**Decoder.** Words from a description, $\{w_t\}_{t=1}^T$, are embedded in the embedding layer: $\hat{w}_{t-1} = \text{Embed}(w_{t-1})$, then sequentially fed to the LSTM layer (Hochreiter and Schmidhuber 1997): $h_t = \text{LSTM}(\hat{w}_{t-1}, h_{t-1})$. We employ the bidirectional attention mechanism (Seo et al. 2017) to align image features and joints features.

$$\tilde{f}^c, \tilde{J}^t, \tilde{f}^t, \tilde{J}^c = \text{CA-Stack}(f^c, \hat{J}^c, f^t, \hat{J}^t) \tag{1}$$

where CA-Stack is a cross attention stack (see Appendix in arxiv full version).

$$f = W_c^\top[\tilde{f}^c; \tilde{f}^t; \tilde{f}^c \odot \tilde{f}^t], \ J = W_c^\top[\tilde{J}^c; \tilde{J}^t; \tilde{J}^c \odot \tilde{J}^t] \tag{2}$$

$$f_t = \text{Att}(h_t, f), \ J_t = \text{Att}(h_t, J), \ J_t^d = \text{Att}(h_t, J^d) \tag{3}$$

$$k_t = W_k^\top[f_t; J_t; h_t; h_t \odot f_t; h_t \odot J_t] \tag{4}$$

$$g_t = W_s^\top[k_t; J_t^d] \tag{5}$$

where Att is general attention (see Appendix in arxiv full version for details). The next token is: $w_t = \text{argmax}(g_t)$, and the loss is: $L_{ML} = -\sum_t \log p(w_t^* | w_{1:t-1}^*, f, J)$, where $w_t^*$ is the GT token.

**RL Training.** We apply the REINFORCE algorithm (Williams 1992) to learn a policy $p_\theta$ upon the model pre-trained with the maximum likelihood approach: $L_{RL} = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)]$; $\nabla_\theta L_{RL} \approx -(r(w^s) - b)\nabla_\theta \log p_\theta(w^s)$, where $w^s$ is a description sampled from the model, $r(\cdot)$ is the reward function, and $b$ is the baseline. We employ the SCST training strategy (Rennie et al. 2017) and use the reward for descriptions from the greedy decoding (i.e., $b = r(w^g)$) as the baseline. We also employ CIDEr as the reward, following Rennie et al. (2017)'s observation (using CIDEr as a reward improves overall metric scores). We follow the mixed loss strategy setup (Wu et al. 2016; Paulus, Xiong, and Socher 2018): $L = \gamma_1 L_{ML} + \gamma_2 L_{RL}$.

**Multilingual Parameter Sharing.** We implement the multilingual training setup by sharing parameters between English and Hindi models, except the parameters of word embeddings, description LSTMs, and final fully connected layers, making the total number of parameters substantially less than those needed for the separate two models summed.

### 6.2 Target Pose Retrieval Model

The current and target candidate images are encoded the same way as the captioning model. A bidirectional LSTM encodes the descriptions: $c = \text{BiLSTM}(\hat{w})$. Image features

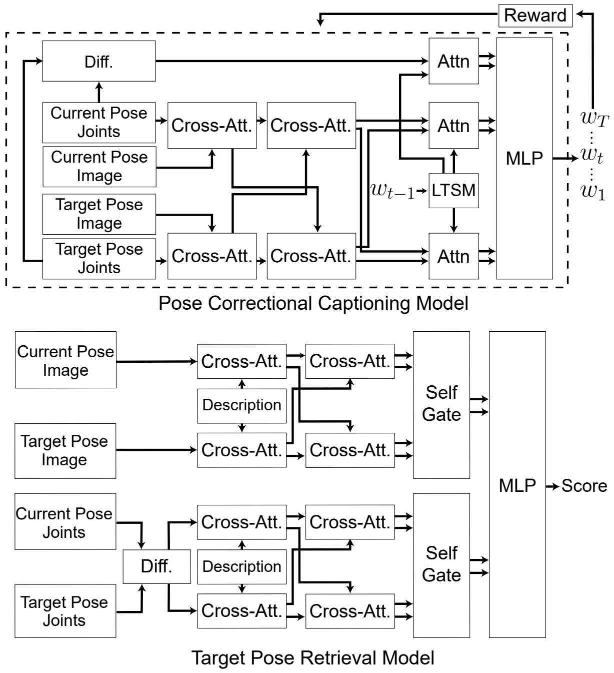


Figure 5: The pose-correctional-captioning model (top) and the target-pose-retrieval model (bottom).

are aligned with description features via cross attention.

$$\tilde{c}^c, \tilde{f}^{t_i}, \tilde{c}^{t_i}, \tilde{f}^c = \text{CA-Stack}(c, f^c, c, f^{t_i}) \tag{6}$$

$$k_{1i} = \text{Self-Gate}([\tilde{c}^c; \tilde{c}^{t_i}; \tilde{c}^c \odot \tilde{c}^{t_i}]) \tag{7}$$

$$g_{1i} = \text{Self-Gate}([\tilde{f}^{t_i}; \tilde{f}^c; \tilde{f}^{t_i} \odot \tilde{f}^c]) \tag{8}$$

where $\odot$ is the element-wise product (see Appendix in arxiv full version for details of the Self-Gate). For joints feature, we calculate the difference between the two joints set: $J^{dt_i} = W_j^\top(J^{t_i} - J^c)$; $J^{dc_i} = W_j^\top(J^c - J^{t_i})$. We apply the same process that the image features go through (i.e., Eq. 6-8) to get $k_{2i}$ and $g_{2i}$.

$$p_i = W_p^\top[k_{1i}; g_{1i}; k_{1i} \odot g_{1i}] \tag{9}$$

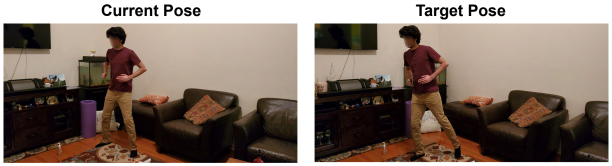
$$q_i = W_q^\top[k_{2i}; g_{2i}; k_{2i} \odot g_{2i}] \tag{10}$$

$$s_i = W_s^\top[p_i; q_i; p_i \odot q_i] \tag{11}$$

The score $s_i$ is calculated for each target candidate and the one with the highest score is considered as the predicted one: $\hat{t} = \text{argmax}([s_0; s_1; ...; s_9])$.

## 7 Experimental Setup

**Data Splits & Training Details.** For the pose-correctional-captioning task, we split the dataset into train/val-seen/val-unseen/test-unseen following Anderson et al. (2018b). We assign separate rooms to val-unseen and test-unseen splits for evaluating model's ability to generalize to unseen environments. The number of task instances for each split is 5,973/562/563/593 (train/val-seen/val-unseen/test-unseen) and the number of descriptions is 5,973/1,686/1,689/1,779. For the target-pose-retrieval task, we split the dataset into train/val-unseen/test-unseen. In this task, "unseen" means "unseen animations". We split the dataset by animations so that the task cannot be easily done by memorizing/capturing patterns of certain animations in the image pairs. After filtering for the target candidates (see Sec. 3), we obtain 4,227/1,184/1,369 (train/val-unseen/test-unseen) instances. We use 512 / 256 as the hidden / word embedding size. We use Adam (Kingma and Ba 2015) as the optimizer. See Appendix in arxiv full version for details.



Description: shift your weight to your right leg. take your left leg off the ground until only your toes are touching the ground. lean you body to the right side and keep your hands on your hip but point your elbows back.

Figure 6: An example from Sim-To-Real transfer dataset.

**Metrics.** For the pose-correctional-captioning task, we employ automatic evaluation metrics: BLEU-4 (Papineni, Roukos et al. 2002), CIDEr (Vedantam, Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005), and ROUGE-L (Lin 2004). Also, motivated by previous efforts towards more reliable evaluation (Wiseman, Shieber, and Rush 2017; Serban et al. 2017; Niu and Bansal 2019; Zhang et al. 2019; Sellam, Das, and Parikh 2020), we introduce new task-specific metrics to capture the important factors. Object-match counts correspondences of environment objects, body-part-match counts common body parts mentioned, and direction-match counts the (body-part, direction) pair match between the model output and the ground-truth (see Appendix in arxiv full version for more information on direction-match). In the target-pose-retrieval task, we use the accuracy of the selection as the performance metric.

**Human Evaluation Setup.** We conduct human evaluation for the pose-correctional-captioning task models to compare the output of the vision-only model, the language-only model, and the full vision+language model qualitatively. We sample 100 descriptions from each model (val-seen split), then asked crowd-workers to vote for the most relevant description in terms of the image pair, and for the one best in fluency/grammar (or 'tied'). Separately, to set the performance upper limit and to verify the effectiveness of our distractor choices for the target-pose-retrieval task, we conduct another human evaluation. We sample 50 instances from the target-pose-retrieval test-unseen split and ask an expert to perform the task for both English and Hindi samples. See Appendix in arxiv full version for more details.

**Unimodal Model Setup.** We implement unimodal models (vision-/language-only) for comparison with the multimodal models. See Appendix in arxiv full version for more details.

**Other Image-Difference Datasets.** We also evaluate our baseline model on other image-difference datasets to show that the baseline is strong and competitive: Image Editing Request (Tan et al. 2019), NLVR2 (Suhr et al. 2019) (the

| Language | Models | Automated Metrics | | | | Task-Specific Metrics | | | Human Eval. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B4 | C | M | R | object-match | body-part-match | direction-match | R | F/G |
| English | V-Only | 6.90 | 6.41 | 16.78 | 30.09 | 0.04 | 1.01 | 0.05 | 4% | 4% |
| | L-Only | 17.74 | 11.42 | 22.14 | 35.16 | 0.08 | 1.22 | 0.15 | 15% | 27% |
| | V+L | 17.55 | 14.47 | 21.29 | 35.21 | 0.18 | 1.29 | 0.13 | 48% | 45% |
| Hindi | V-Only | 8.43 | 4.37 | 18.90 | 28.55 | 0.03 | 1.21 | 0.02 | 9% | 10% |
| | L-Only | 25.42 | 11.41 | 29.68 | 36.90 | 0.0 | 1.42 | 0.07 | 19% | 26% |
| | V+L | 18.99 | 8.58 | 29.26 | 34.73 | 0.08 | 1.63 | 0.10 | 51% | 53% |

Table 2: The performance of the unimodal and multimodal models on automated metrics, our new task-specific metrics, and human evaluation. for both English and Hindi dataset on the val-seen split (B4: BLEU-4, C: CIDEr, M: METEOR, R: ROUGE, V: Vision+Joints, L: Language, R: Relevancy, F/G: Fluency and Grammar).

| Language | Models | B4 | C | M | R |
|---|---|---|---|---|---|
| English | V+L | 17.55 | 14.47 | 21.29 | 35.21 |
| | (-) Joints | 17.39 | 13.79 | 21.35 | 34.86 |
| | (+) RL | 18.69 | 16.04 | 22.35 | 36.18 |
| | (+) Multi-L | 19.08 | 15.71 | 22.47 | 36.46 |
| Hindi | V+L | 18.99 | 8.58 | 29.26 | 34.73 |
| | (-) Joints | 18.23 | 7.93 | 27.55 | 34.12 |
| | (+) RL | 18.57 | 9.63 | 28.83 | 34.76 |
| | (+) Multi-L | 18.67 | 9.77 | 29.05 | 34.74 |

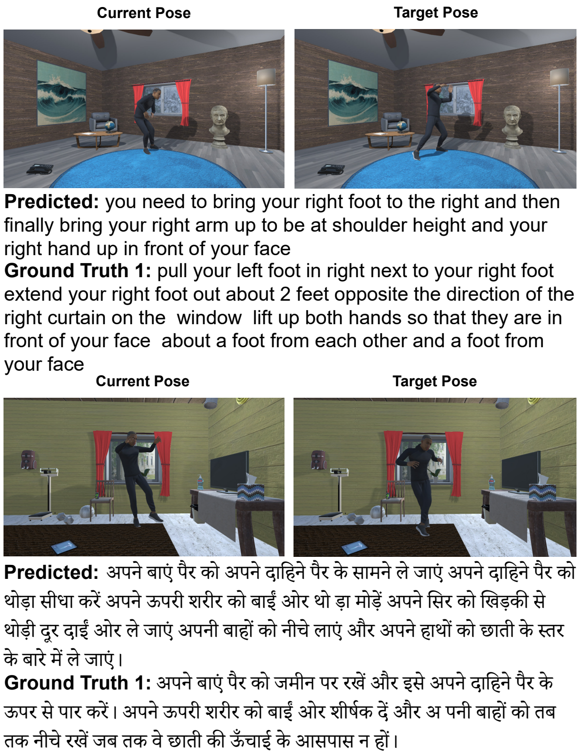
Table 3: Model ablations on val-seen split (RL: reinforcement learning, Multi-L: multilingual).

| Dataset | Model | B4 | C | M | R |
|---|---|---|---|---|---|
| Image Editing Request<br>Tan et al. (2019) | DRA | 6.72 | 26.36 | **12.80** | 37.25 |
| | Ours | **7.88** | **27.70** | 12.53 | **37.56** |
| NLVR2<br>Suhr et al. (2019) | DRA | 5.00 | **46.41** | 10.37 | **22.94** |
| | Ours | **5.30** | 45.09 | **10.53** | 22.79 |
| CLEVR-Change (SC)<br>Park et al. (2019) | DUDA | 42.9 | 94.6 | 29.7 | - |
| | Ours | **44.0** | **98.7** | **33.4** | 65.5 |

Table 4: Our baseline V+L model performs competitively on other image-difference captioning datasets (DRA: Dynamic Relation Attention (Tan et al. 2019), DUDA: Dual Dynamic Attention Model (Park, Darrell, and Rohrbach 2019); SC = Scene Change).

variant from Tan et al. (2019)), and CLEVR-Change (Park, Darrell, and Rohrbach 2019).

**Sim-to-Real Transfer.** To verify the possibility of the transfer of our simulated image dataset to real images, we collect real image pairs of current and target poses. We randomly sample 60 instances from test-unseen split (test-sim) and then the authors and their family members[4] follow the poses in the sampled test-sim split to create the real image version (test-real). Since the environments (thus objects and their layout too) and poses (though they are told to try to match as accurately as possible) have differences between the two splits (i.e., test-sim and test-real), we manually rewrite a few words or phrases in the descriptions to make it more consistent with images in the test-real split (see Fig. 6).



**Predicted:** you need to bring your right foot to the right and then finally bring your right arm up to be at shoulder height and your right hand up in front of your face

**Ground Truth 1:** pull your left foot in right next to your right foot extend your right foot out about 2 feet opposite the direction of the right curtain on the  window  lift up both hands so that they are in front of your face  about a foot from each other and a foot from your face

**Predicted:** अपने बाएं पैर को अपने दाहिने पैर के सामने ले जाएं अपने दाहिने पैर को थोड़ा सीधा करें अपने ऊपरी शरीर को बाईं ओर थो ड़ा मोड़ें अपने सिर को खिड़की से थोड़ी दूर दाईं ओर ले जाएं अपनी बाहों को नीचे लाएं और अपने हाथों को छाती के स्तर के बारे में ले जाएं।

**Ground Truth 1:** अपने बाएं पैर को जमीन पर रखें और इसे अपने दाहिने पैर के ऊपर से पार करें। अपने ऊपरी शरीर को बाईं ओर शीर्षक दें और अ पनी बाहों को तब तक नीचे रखें जब तक वे छाती की ऊँचाई के आसपास न हों।

Figure 7: Output examples of our multimodal model in English (top) and Hindi (bottom); only showing 1 GT due to space limitations.

# 8 Results

## 8.1 Pose Correctional Captioning Task

As shown in Table 2, the V+L models show better performance than V-only models. The L-only model shows higher scores on some of the automatic metrics, likely because the descriptions in our FixMyPose dataset are instructional about body parts (and their movements/directions), so similar phrases are repeated and shallow metrics will only focus on such phrase-matching, not correctly reflecting human evaluations (Belz and Reiter 2006; Reiter and Belz 2009; Scott and Moore 2007; Novikova et al. 2017; Reiter 2018). Thus, we also evaluate the output of each model on our task-specific metrics that account the important factors (ob-

[4] Hence covering diverse demographics, including some that are different from the simulator data splits, as well as different room environments. All participants consented to the collection of images (and additionally, we blur all faces).

| Split | Automated Metrics |  |  |  | Task-Specific Metrics |  |
|---|---|---|---|---|---|---|
|  | B4 | C | M | R | OM | DM |
| test-sim | 16.93 | 9.91 | 21.79 | 35.08 | 0.04 | 0.20 |
| test-real | 13.01 | 7.12 | 21.40 | 33.05 | 0.07 | 0.11 |

Table 5: Sim-to-Real transfer performance. Since there is no GT joints for real images, the body-part-match metric is not available (OM: object-match, DM: direction-match).

jects, body parts, and movement directions), and we also conduct human evaluation to check the real quality of the outputs. The V+L models show better performance on the task-specific metrics and human evaluation, meaning they capture essential information and their outputs are more relevant to the images and more fluent in the respective language. See Appendix for "unseen" split results.[5]

**Ablations.** As Table 3 shows, adding body joints features improves the score much, implying body joints gives additional important information to capture human movements.

**RL/Multilingual Model Results.** As Table 3 shows, RL training helps improve scores by directly using the evaluation metric (CIDEr) as the reward. We leave exploring more effective reward functions (e.g., the joints distance from a reverse pose generation task) for future work. Table 3 also shows that the multilingual training setup achieves comparable scores (similar observation to Wang et al. (2019)) with only 71% of the parameters of the separate training setup (13.2M vs 18.7M), promising future work on more compact and efficient multilingual models.

**Other Image-Difference Datasets.** Table 4 shows that our V+L baseline model beats or matches state-of-the-art models on other datasets, implying our baseline models are strong starting points for our FIXMYPOSE dataset.

**Output Examples.** Outputs from our V+L models are presented in Fig. 7. The English model captures the movement of the character's legs and arms ("bring your right foot to the right" and "bring your right arm up to be at shoulder height ... right hand up in front of your face"). The Hindi model captures movement of the body parts and their spatial relationship to each other (English translation: "move your left leg in front of your right leg..."), the model can also describe movement using object referring expressions (English translation: "...move your head slightly away from the window..."). See Fig. 7 for the original Hindi and Appendix in arxiv full version for full analysis and unimodal outputs.

**Demographic Ablations.** We evaluate our V+L model on individual character avatar. The results show our dataset is not skewed to favor a specific demographic or character. Please see the detailed scores in arxiv full version.

**Sim-to-Real Transfer.** As shown in Table 5, the sim-to-real performance drop is not large, meaning information learned from our simulated FIXMYPOSE dataset can be transferred to real images reasonably well. Also, considering that the results are from a set of images of people with different demographics and different environments, there is no particular bias in the models' output which is trained on our dataset. Since there is no GT body joints for the real images, we modify our model so it can also be trained to predict the joints during training time as well as generate descriptions (multi-task setup) and use the estimated joints at test time.[6]

| Models | Accuracy (%) |  |
|---|---|---|
|  | English | Hindi |
| Random-Selection | 9.81 |  |
| V-Only | 34.82 |  |
| L-Only | 8.86 | 8.96 |
| V+L | 38.49 | 37.84 |
| Human | 96.00 | 96.00 |

Table 6: The scores for the target-pose-retrieval task. While the V+L models scores the highest, there is still much room for improvement when compared with human performance.

### 8.2 Target Pose Retrieval Task

As shown in Table 6, V+L models show the highest scores for the target-pose-retrieval task, indicating that achieving high performance is not possible by exploiting unimodal biases. V-Only models score higher than the random-selection model, which selects an image at random, because even with our careful distractor choices (see Sec. 3 and Appendix in arxiv full version), the poses in the "current" and "target" images are more similar to each other than the other images. However, the human-model performance gap is still quite large, implying there is much room for improvement.[7]

## 9 Conclusion and Future Work

We introduced FIXMYPOSE, a novel pose correctional description dataset in both English and Hindi. Next, we proposed two tasks on the dataset, pose-correctional-captioning and target-pose-retrieval, both of which require models to understand diverse linguistic properties such as egocentric relation, environmental direction, implicit movement description, and analogous reference as well as capture fine visual movement presented in two images. We also presented unimodal and multimodal baselines as strong starter models.Finally, we demonstrated the possibility of transfer to real images. In future work, we plan to further expand the FIXMYPOSE dataset with more languages and even more diversity in the character pool (e.g., height, age, etc. based on digital avatar availability) and animations.

[5] We also checked for variance by running models with 3 different seeds and the stddev is small (less than/near 0.5% on CIDEr).

[6] For the simulated data results in Table 3 (English), we obtain a CIDEr score of 14.17 using predicted joints (on the val-seen split), which as expected is between the non-joint (13.79) and GT-joint (14.47) models' results (hence showing that reasonable performance can be achieved without GT joints at test time). The average distance between predicted and GT joints is around 0.4 meters.

[7] Human performance is 96% when given the full task (English), but much lower when only given lang. (38%) or only vis. (22%), further indicating that both lang.+vis. is needed to solve the task.

# Acknowledgments

We thank the reviewers and Jason Baldridge, Peter Hase, Hao Tan, and other UNC-NLP members for their helpful comments. This work was supported by NSF Award 1840131, NSF-CAREER Award 1846185, DARPA MCS Grant N66001-19-2-4031, Microsoft Investigator Fellowship, and Google Focused Award. The views contained in this article are those of the authors and not of the funding agency.

# Ethics Statement

Our paper and dataset hopes to enable people to improve their health and well-being, as well as strives to follow ethical standards, e.g., we especially try to maintain balance across diverse demographics and avoid privacy concerns by collecting data from a simulated environment (but still show good transfer to real images), and we also expand beyond English so as to more inclusively cover multiple languages. Similar to other image captioning tasks/models, some imperfect descriptions from models trained on our FixMyPose dataset might also lead to difficult/unnatural poses. Presenting models' confidence scores can help people ignore such unnatural pose corrections; however, most importantly, careful use is required for real-world applications (similar to all other image captioning models/tasks, e.g., the ones used for accessibility and visual assistance), and further broader discussion on developing fail-safe AI systems is needed.

# References

Abbasi, B.; Monaikul, N.; Rysbek, Z.; Di Eugenio, B.; and Žefran, M. 2019. A Multimodal Human-Robot Interaction Manager for Assistive Robots. In *IROS*, 6756–6762. IEEE.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; et al. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 3674–3683.

Andriluka; Pishchulin; Gehler; and Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *CVPR*.

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; et al. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, 65–72.

Belz, A.; and Reiter, E. 2006. Comparing automatic and human evaluation of NLG systems. In *EACL*.

Bisk; Shih; Choi; and Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *AAAI*.

Bisk, Y.; Marcu, D.; and Wong, W. 2016. Towards a dataset for human computer communication via grounded language acquisition. In *Workshops at AAAI*.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *TPAMI* .

Forbes, M.; Kaeser-Chen, C.; Sharma, P.; and Belongie, S. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *EMNLP*. Hong Kong.

Gehring; Auli; Grangier; Yarats; and Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR* 47: 853–899.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *EMNLP*, 4024–4034.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 4565–4574.

Johnson, S.; and Everingham, M. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*. Doi:10.5244/C.24.12.

Johnson, S.; and Everingham, M. 2011. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 1465–1472. IEEE.

Kim, K.; Boelling, L.; Haesler, S.; Bailenson, J.; et al. 2018. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In *ISMAR*. IEEE.

Kim, K.; de Melo, C. M.; Norouzi, N.; Bruder, G.; and Welch, G. F. 2020. Reducing Task Load with an Embodied Intelligent Virtual Assistant for Improved Performance in Collaborative Decision Making. In *2020 IEEE VR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 317–325.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* .

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; et al. 2019. Dancing to Music. In *NeurIPS*.

Lei; Yu; Berg; and Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*.

Li, S.; Scalise, R.; Admoni, H.; et al. 2016. Spatial references and perspective in natural language instructions for collaborative manipulation. In *RO-MAN*. IEEE.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 375–383.

Melas-Kyriazi, L.; Rush, A. M.; and Han, G. 2018. Training for diversity in image paragraph captioning. In *EMNLP*.

Niu, T.; and Bansal, M. 2019. Automatically Learning Data Augmentation Policies for Dialogue Tasks. In *EMNLP*.

Novikova; Dušek; Curry; and Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *EMNLP*.

Papineni, K.; Roukos, S.; et al. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *ICCV*, 4624–4633.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A Deep Reinforced Model for Abstractive Summarization. In *ICLR*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2641–2649.

Reiter, E. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics* 44(3): 393–401.

Reiter, E.; and Belz, A. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* 35(4).

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; et al. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Rong, Y.; Shiratori, T.; and Joo, H. 2020. FrankMocap: Fast Monocular 3D Hand and Body Motion Capture by Regression and Integration. *arXiv preprint arXiv:2008.08324* .

Saunders, B.; Camgoz, N. C.; and Bowden, R. 2020. Progressive Transformers for End-to-End Sign Language Production. *ECCV 2020* .

Scott, D.; and Moore, J. 2007. An NLG evaluation competition? eight reasons to be cautious. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, 22–23.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *ACL*.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional Attention Flow for Machine Comprehension. In *ICLR*.

Serban; Sordoni; Lowe; Charlin; Pineau; Courville; and Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*.

Shlizerman, E.; Dery, L.; Schoen, H.; and Kemelmacher-Shlizerman, I. 2018. Audio to body dynamics. In *CVPR*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *ACL*, 6418–6428.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 5693–5703.

Tan; Dernoncourt; Lin; Bui; and Bansal. 2019. Expressing Visual Relationships via Language. In *ACL*.

Tang, T.; Jia, J.; and Mao, H. 2018. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *ACM Multimedia*.

Toshev, A.; and Szegedy, C. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Vedantam; Zitnick, L.; and Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Verma, M.; Kumawat, S.; Nakashima, Y.; and Raman, S. 2020. Yoga-82: a new dataset for fine-grained classification of human poses. In *CVPR Workshops*, 1038–1039.

Wang, I.; Smith, J.; and Ruiz, J. 2019. Exploring virtual agents for augmented reality. In *CHI*, 1–12.

Wang, S. I.; Liang, P.; and Manning, C. D. 2016. Learning Language Games through Interaction. In *ACL*, 2368–2378.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional Pose Machines. In *CVPR*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4): 229–256.

Wiseman, S.; Shieber, S. M.; and Rush, A. M. 2017. Challenges in Data-to-Document Generation. In *EMNLP*.

Wu; Schuster; Chen; Le; Norouzi; Macherey; Krikun; Cao; Gao; Macherey; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .

Xu; Ba; Kiros; Cho; Courville; Salakhudinov; Zemel; and Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI* .

Yang; Yuan; Wu; Cohen; and Salakhutdinov. 2016. Review networks for caption generation. In *NeurIPS*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *ECCV*, 684–699.

Zhang, T.; Kishore, V.; Wu, F.; et al. 2019. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Zhao, L.; Peng, X.; et al. 2019. Semantic graph convolutional networks for 3D human pose regression. In *CVPR*.

Zhuang, W.; Wang, Y.; Robinson, J.; Wang, C.; Shao, M.; Fu, Y.; and Xia, S. 2020. Towards 3D Dance Motion Synthesis and Control. *arXiv preprint arXiv:2006.05743* .