CHORUS: Learning Canonicalized 3D Human-Object Spatial Relations from Unbounded Synthesized Images (Supplementary Material)

A. Additional Details for Method

In this section, we provide further details on the formulations and implementations regarding the method section (Sec. 3) of our main paper.

A.1. Prompt Generation

We generate 3 to 20 different semantic HOI prompts via chatGPT [51] using the following query template, where {**m**} is replaced with an integer between 3∼20 and {**category**} is replaced with one of our input category keywords:

> *"Generate at most* {**m**} *simple subject-verb-object prompt where subject's* {category is person/word is "A person"} *and object's category is* {**category**}. *You should use diverse and general word but no pronoun for subject. Generated prompt must align with common sense. Verb must depict physical interaction between subject and object. Simple verb preferred."*

The categories considered in this work are listed in Tab. 3. After generating HOI prompts, we augment each prompt with $m_{\text{aug}} = 22$ different viewpoint augmentations, including:

> *(no augmentation) / side view / back view / front view / top view / bottom view / realistic photo / seen from side / seen from back / seen from various views / seen from close view / seen from far away / scenic view / full body photo / hand photo / face photo / photo taken close to hand / photo taken close to face / selfie / close-up photo / hand only / face only*.

A.2. Synthesizing Text-Conditioned Images via Diffusion

We use publicly available Stable-Diffusion [59] model (version: 1.4) for text-to-image synthesis. We generate 512×512 images upsampled from the generated $8\times$ downsampled latents. We sample images with 50 denoising steps using Pseudo-Numerical-Sampling methods [40] with classifier-free guidance scale of 7.5.

A.3. Filtering

We apply a cascaded framework to remove unrelated images. We use PointRend [37] for bounding-box detection and instance segmentation of COCO [39] categories. For LVIS [20] categories (e.g., hat, sunglasses, sweater), we use publicly available pretrained Mask-RCNN [24] model from detectron2 [76] where we set the segmentation threshold as 0.8. We post-process the predicted instances by removing duplicated bounding boxes on the same target object. Specifically, we remove the lower-confidence instance if two bounding boxes of the same category overlap with the *intersection over smaller bounding box* value bigger than 0.8, with exceptions of bounding boxes with confidence over 0.98. We also filter the images with multiple or none human/object (target category) instances. We then reject images when the *intersection over object bounding box* value between the human box and object box is smaller than 0.1, assuming there is no interaction between them. For keypoint-filtering, we represent the keypoints in COCO [39] format. We use a top-down approach using publicly available pretrained HRNet [70]+Darkpose [82] (provided by MMPose [13]) with a confidence threshold of 0.7 for keypoint prediction. We exclude the image if no shoulder joints (i.e., *left-shoulder, right-shoulder*) or no hip joints (i.e., *left-hip, right-hip*) exist. Finally, we reject the images with a very small human bounding box that returns no prediction from 3D human pose estimator [60].

Table 3. **Statistics for generated dataset for all categories.**

Category	# Prompts	# Images	# Images after Filtering	Rejection-rate (%)
Motorcycle	12	47520	18228	61.64
Bench	12	31680	15193	52.04
Backpack	20	42240	7530	82.17
Handbag	4	23760	1374	94.22
Tie	10	31680	3358	89.40
Frisbee	3	21780	5502	74.74
Skis	12	34848	14547	58.26
Snowboard	4	7920	3251	58.95
Sports ball	8	19008	1611	91.52
Baseball glove	5	11880	260	97.81
Skateboard	16	50688	12069	76.19
Surfboard	16	73920	25352	65.70
Tennis racket	15	47520	19827	58.28
Cell phone	20	84480	2344	97.23
Bicycle	18	42768	11361	73.44
Umbrella	7	16632	2720	83.65
Chair	17	47124	10185	78.39
Bed	5	11880	1856	84.38
Laptop	17	53856	1890	96.49
Hat*	10	31680	173	99.45
Sweater*	3	5940	689	88.40
Sunglasses*	3	5940	78	98.69
Soccer ball*	5	15840	1976	87.53
Scarf*	5	15840	461	97.09

* denotes LVIS [20] categories

A.4. Viewpoint Estimation via 3D Human Pose Estimation

While the weak perspective is sufficient for projection, we employ a perspective camera to consider the distance

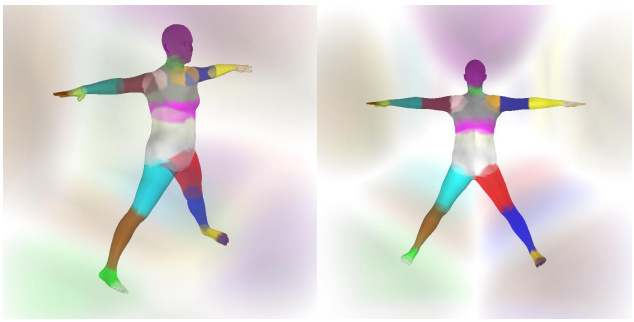

Figure 13. **Visualization of LBS weights.**

between the camera and the human subject. To convert the weak perspective camera model π and camera-centric orientation ϕ (refer to Eq. 3) into a perspective camera model Π in the person-centric coordinate system (i.e., the origin is defined at the pelvis), we apply optimization to compute Π by aligning 3D joint projections and $\mathbf{j}$:

$$\Pi^* = \arg\min_{\Pi} \sum_i ||\Pi(\mathbf{J}_i^0) - \mathbf{j}_i||^2, \tag{8}$$

where i is the joint index and $\mathbf{J}^0$ are the 3D joints of the SMPL model in person-centric coordinate, which is simply obtained by putting zero orientation for ϕ while keeping other parameters θ and β. We include the joints $\mathbf{j}_i$ outside the image range for optimization described in Eq. 8. We optimize the joint-reprojection loss using Adam [36] with a learning rate of 0.01 for 2400 iterations. We early-terminate if the joint-reprojection loss is below 0.7 in pixel scale. We initialize camera intrinsic parameters with field-of-view $46.4°$ and camera extrinsic parameters with the rotation matrix achieved by Rodrigues formula using ϕ and the translation vector by $[t_x, t_y, 2f/s]$ where t_x, t_y, s are each x,y-direction translations, and scale from weak-perspective camera π.

A.5. 3D Occupancy Estimation via Human Pose Canonicalization

We define LBS weights for arbitrary point $\mathbf{x}^c$ in canonical space following Eq. 4 with $\mathbf{k} = 30$. We encourage "zero motions" by mixing computed LBS weights with standard basis vector for 1st dimension $\mathbf{e}_1 \in \mathbb{R}^{24}$:

$$\omega_{\text{deweight}}(\mathbf{x}^c) = (1-\alpha)\mathbf{e}_1 + \alpha\omega(\mathbf{x}^c) \tag{9}$$

$$\alpha = |\max(\frac{\tau - d_{\text{min}}}{\tau + d_{\text{min}}}, 0)|^s \tag{10}$$

where τ and s are each bandwidth and smoothing hyperparameters we choose as $\tau = 0.8, s = 0.25$, and d_{min} is the distance between $\mathbf{x}^c$ and nearest SMPL mesh vertex. Note that the 1st element of LBS weight is related to the pelvis joint, which is fixed to the origin and aligned in a person-centric coordinate system as we set $\phi = 0$; hence, the transformation matrix is identity in $\text{SE}(3)$. Finally, we apply Laplace smoothing over the entire grid 30 times to smooth LBS weights, similar to SelfRecon [30]. See Fig. 13 for visualization of extended LBS weights.

Table 4. **Definition of Body Parts.** Each body part is defined by merging multiple SMPL body segmentation maps from Meshcapade [45].

Body Part	SMPL Body Segmentation Labels
rightHand	rightHand, rightHandIndex1
leftHand	leftHand, leftHandIndex1
rightArm	rightArm, rightForeArm
leftArm	leftArm, leftForeArm
rightLowerLeg	rightLeg
leftLowerLeg	leftLeg
rightUpperLeg	rightUpLeg
leftUpperLeg	leftUpLeg
rightFoot	rightFoot, rightToeBase
leftFoot	leftFoot, leftToeBase
torso	spine, spine1, spine2, leftShoulder, rightShoulder
face	head, neck

A.6. Uniform View Sampling

Before performing aggregation, we first check the camera distribution and assign the accumulation score r_k for each image. Specifically, we divide the azimuth region $[0, 2\pi)$ into 12 equispaced bins (each $\frac{\pi}{6}$ long) and count the number of cameras in each bin. If the camera associated with the image falls into a specific bin, r_k is set as the inverse of the camera numbers in that bin. After assigning r_k for all images, we then perform the aggregation following Eq. 6 or Eq. 13.

A.7. Inference for Posed Space

At inference, we first calculate LBS weights that transform the voxels from pose-deformed space to canonical space in the same way of Eq. 4 and Eq. 9, where in this case $\mathbf{x}^c$ is replaced with $\mathbf{x}$ and $\mathbf{v}_i$ corresponds to location of i-th vertex in SMPL in pose-deformed space. Denoting the j-th LBS weights in pose-deformed space as $\omega_j^{\text{inv}}(\mathbf{x}; \theta)$, we inversely deform the voxels in pose-deformed space as:

$$\mathbf{x}^c = \mathcal{W}^{-1}(\mathbf{x}) = \sum_{j=1}^{n_b} \omega_j^{\text{inv}}(\mathbf{x}; \theta) \cdot \mathbf{B}_j(\theta_j)^{-1} \cdot \mathbf{x} \tag{11}$$

We set $\Phi_o(\mathbf{x}|\theta)$ as the learned occupancy $\Phi_o^c(\mathbf{x}^c)$ in canonical distribution to infer pose-deformed distribution.

A.8. Selective Aggregation via Semantic Clustering

We note that body part $\mathbf{a} \in \mathbf{A}$ is a hyperparameter that can be easily given by annotating a set of SMPL mesh vertices with a body part label. In practice, we define $\mathbf{A}$ as 12 body parts obtained by merging publicly available SMPL segmentation maps from Meshcapade [45]. Correspondence between defined body parts and SMPL body segmentation labels is provided in Tab. 4. Representing each body part $\mathbf{a} \in \mathbf{A}$ as a binary operator that outputs 1 if the corresponding SMPL mesh vertex is part of the body part $\mathbf{a}$ else 0, we

can define the interaction region for $\mathbf{a}$ in the canonical space as below:

$$\mathbf{I_a} = \bigcup_{i;\ \mathbf{a}(\mathbf{v}^i)=1} \{\mathbf{x}^c | \ ||\mathbf{x^c} - \mathbf{v}_i|| \leq \epsilon\} \tag{12}$$

where ϵ is the interaction threshold, where we set as $\epsilon = 0.13$. The interaction region $\mathbf{I_a}$ plays a role of determining whether the provided image contains the HOI involving contact with body part $\mathbf{a}$. Specifically, we ignore the image if none of the 3D canonical points within $\mathbf{I_a}$ are warped and projected into the object mask. Putting it all together, selective aggregation for 3D occupancy can be described as:

$$\Phi_o^c(\mathbf{x^c}|s) = \frac{\sum_{k=1}^{|\mathbf{G}(p)|} r_k \mathcal{M}_k(\Pi_k(\mathcal{W}(\mathbf{x^c}))) \cdot \mathbf{1}(1 \in \mathcal{M}_k(\Pi_k(\mathcal{W}(\mathbf{I_a}))))}{\sum_{k=1}^{|\mathbf{G}(p)|} r_k \mathcal{I}_k(\Pi_k(\mathcal{W}(\mathbf{x^c}))) \cdot \mathbf{1}(1 \in \mathcal{M}_k(\Pi_k(\mathcal{W}(\mathbf{I_a}))))} \tag{13}$$

where $\mathbf{G}(p)$ denotes the set of generated images from single HOI prompt $p \in \mathbf{P}$ and $\mathbf{1}(\cdot)$ is a binary operator that returns 1 if the provided input is true else 0. At inference, learned occupancy probabilities are used to compute $\Phi_o(\mathbf{x}|\theta, s)$ (refer to Eq. 1) as described in Supp. Mat. A.7.

B. Additional Details for Experiments

B.1. Dataset

Generated Dataset. We refer readers to Tab. 3 for full per-category statistics for the generated dataset.

Image Search Dataset. We use $m = 12$ prompts (same as the prompts used for generating dataset) for category *motorcycle* and $m_{\text{aug}} = 22$ viewpoint augmentations (same as in Supp. Mat. A.1) for image search, and we set the desired number of retrieved images as $N = 1000$ or $N = 10000$. We crawl up to $\lfloor \tau_{\text{mult}} \times \frac{N}{m \times m_{\text{aug}}} \rfloor + \tau_{\text{add}}$ image links, where τ_{mult} and τ_{add} are introduced to tolerate the number of undownloadable/unreadable files. Specifically, we set $\tau_{\text{mult}} = 1.1$ and $\tau_{\text{add}} = 1$. Subsequently, collected images are resized to a shorter side length of 512 and center-cropped, resulting in 512×512 images.

Extended COCO-EFT Dataset for Testing. We provide detailed dataset preparation procedures for the COCO-EFT [32] dataset. Similar to filtering (Sec. 3.2), we only retain samples with a single human, single object (of target category), and filter the images based on *intersection over smaller bounding box* value between the human and the object. We assume no human-object interaction if this value is below 0.5. It is important to note that we do not apply any manual filtering to ensure fairness in our evaluation. After filtering, we compute the perspective camera parameters following Supp. Mat. A.4, except we optimize for 3000 iterations and early-terminate if the joint-reprojection loss is below 0.5 in pixel scale. We reject the image if the joint-reprojection loss is over 1.0 in pixel scale. To ensure multi-viewpoint evaluation, we only use categories with more than 30 images in the extended COCO-EFT dataset. Refer to Tab. 5 for the summary of statistics.

Table 5. **Statistics in the Extended COCO-EFT Dataset.** We report number of images for categories with a minimum of 30 images in COCO-EFT [32] dataset.

Category	Number of Images in Dataset
Motorcycle	36
Bench	37
Backpack	83
Handbag	47
Tie	37
Frisbee	36
Skis	86
Snowboard	67
Sports ball	67
Baseball glove	49
Skateboard	176
Surfboard	110
Tennis racket	117
Cell phone	60

B.2. Projective Average Precision

We provide detailed protocols for PAP evaluation and intuition behind each step in this section. Briefly speaking, the PAP metric quantifies the validity of object occupancy distribution in pose-deformed 3D space (current space) without 3D annotations by comparing the projection of distribution with 2D annotation from multiple viewpoints. Given the category keyword to evaluate, we first start by deforming the distribution from canonical space to pose-deformed space using annotated SMPL pose from the test dataset following the inference method described in Sec. 3.3. Note that we can get probability occupancy values for equispaced gridpoints in pose-deformed space. We discretize the distribution in pose-deformed space for various threshold values and project binary occupancy using an annotated perspective camera. Discretization is applied to bypass the ambiguity of mixing probabilities when more than one 3D probability value falls into the same pixel in 2D. We use all thresholds from $0.01 \sim 1$ equispaced with interval 0.01. Using multiple thresholds enables us to evaluate the distribution regarding the intra-class variation of object geometry. Next, we compute *pixel-wise* precision and recall between rendered mask and annotated object segmentation mask for all thresholds. Specifically, we downsample rendered mask and object segmentation mask preserving aspect ratio with a shorter length being 32 before we compute precision and recall. We downsample for two reasons: (1) to allow other 3D representations that require high-compute in the PAP evaluation pipeline, and (2) to tolerate the variance in object size and geometry. Note that the goal of the PAP metric is to evaluate *validity* of the distribution, not the accuracy or quality of the reconstruc-

tion. Finally, we compute interpolated AP similar to Pascal VOC 2008 [17] using precision-recall values, which are subsequently averaged across all test images within the given category to yield the PAP value. We report two different PAP metrics in terms of interpolation methods when predicted occupancy is entirely 0 after discretization:

- *Vanilla*: Uses the highest precision value from lower thresholds when discretized distribution is entirely 0
- *Strict*: Sets precision value as 0 when discretized distribution is entirely 0

Note that Human-Occlusion-Aware PAP metrics follow the same protocols, except we exclude the human-occluded region when computing precision and recall. We do not apply semantic clustering during quantitative evaluation, i.e., we evaluate marginalized distribution aggregated with all images generated from all prompts per category.

C. Additional Qualitative Results

We report additional qualitative results for various categories in this section. Same as in Sec. 4.4, we use SMPL pose sampled from the extended COCO-EFT dataset for COCO categories or generated dataset for LVIS categories to deform the distribution in canonical space to pose-deformed space. See Fig. 14 $\sim$ 36 for results.

D. Limitations & Future Works

Granularity. Our method returns a plausible set of object distributions; however, we represent them as a low-resolution voxel field (resolution 48^3), which limits the representability and granularity of the results. Future research can explore alternative 3D representations (e.g., volume rendering-based methods similar to NeRF [46]) to improve computation efficiency and achieve higher quality.

Problems with Small Objects. Our method particularly shows weakness in small objects, especially those interacting with hands, primarily due to heavy occlusion and expressivity constraints in the SMPL model. For future research, employing the SMPL-X [53] representation to learn distributions for small objects interacting with hands, or using close-view cameras from various angles to reduce occlusion, could be beneficial.

Bias and Artifacts in Synthesized Images. We use viewpoint augmentation to control the camera distribution during image synthesis; however, this method lacks full controllability and requires improvement. Although we minimize this effect by assigning camera distribution-aware accumulation scores, there is still a possibility of bias. One potential approach to address this challenge is by employing the PerpNeg algorithm [1] to enhance viewpoint control during generation. Additionally, synthesized images are likely to contain artifacts, which could propagate errors in later steps (e.g., human prediction) and lead to incorrect modeling of the occupancy probability distribution. Improving image synthesis methods will help mitigate such challenges.

Bias due to Heavy Filtering. As our filtering strategy involves various off-the-shelf methods, employing heavy filtering may introduce bias. For example, the object detection method we use may be imperfect and could filter out images even if an object is present, leading to bias in the generated dataset after filtering. Consequently, this may result in bias in the occupancy probability distribution. Soft filtering methods (i.e., applying confidence weights to each image instead of removing images with hard thresholds) may be an alternative, which we leave for future work.

Modelling Multimodal Scenarios. The semantic clustering step in our method provides a means to understand objects that humans can interact with in various ways. While our method effectively models plausible HOI exhibiting specific interaction types, it requires manual definition of body parts and specification of HOI prompts to represent the semantics. Additionally, relying on user evaluation for identifying plausible semantic clusters hinders the efficient and automatic expansion of the corpus, as this process becomes manual. We acknowledge the need for further research to enhance the automatic generation of 3D HOI spatial relations without this manual constraint.

Category Limits. Currently, our method mainly considers categories from COCO [39] and a few categories from LVIS [20] due to the availability of the object detection and segmentation method. A promising future direction is to replace this current object detector with open-vocabulary models like ODISE [78] to incorporate additional categories.

Evaluation Metrics. Our evaluation metric (PAP) has room for improvement. For example, our current method utilizes a simple downsampling strategy to smooth the distribution during the protocol, which could be enhanced with other strategies (e.g., kernel-smoothing methods). Additionally, it is worth exploring enhancements to the metric that can effectively quantify the validity of multimodal distributions.

Potential Downstream Applications. Our method exhibits myriads of potential downstream applications, such as; (1) using our extracted knowledge as prior for HOI modeling (e.g., replacing interaction labels in PHOSA [83]); (2) improving action recognition methods based on the current pose of the human and object; (3) scene generation and object localization from human postures; or (4) applications for robotics.

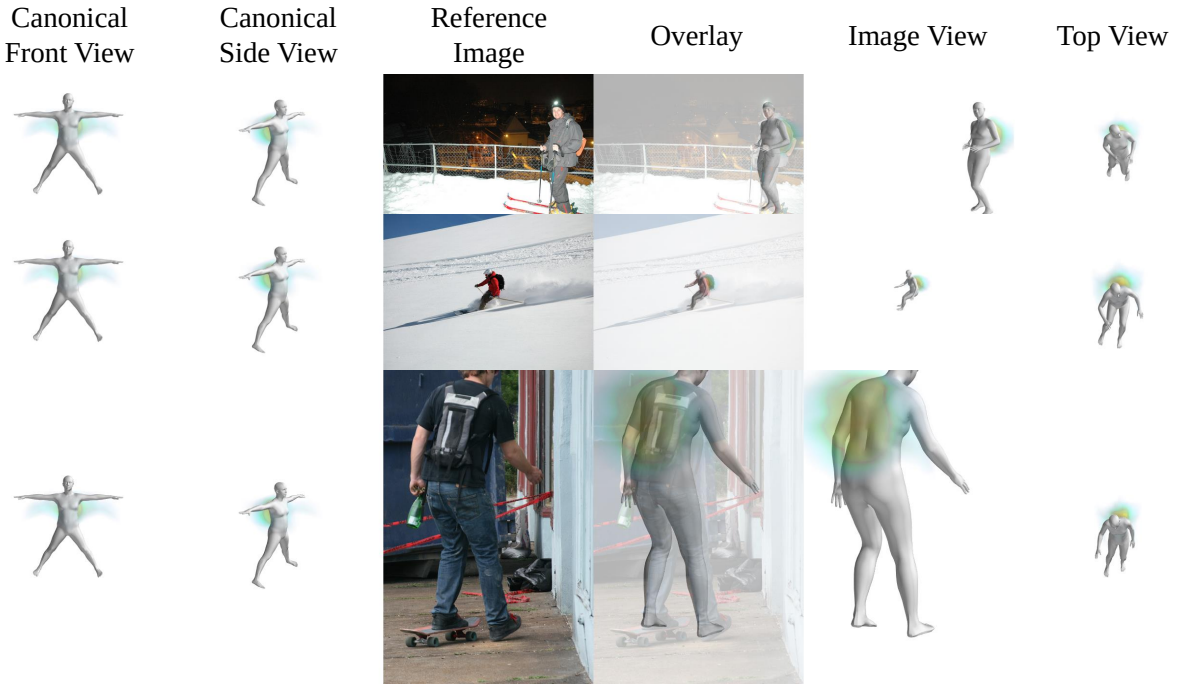

Figure 14. Qualitative results for category **backpack**.

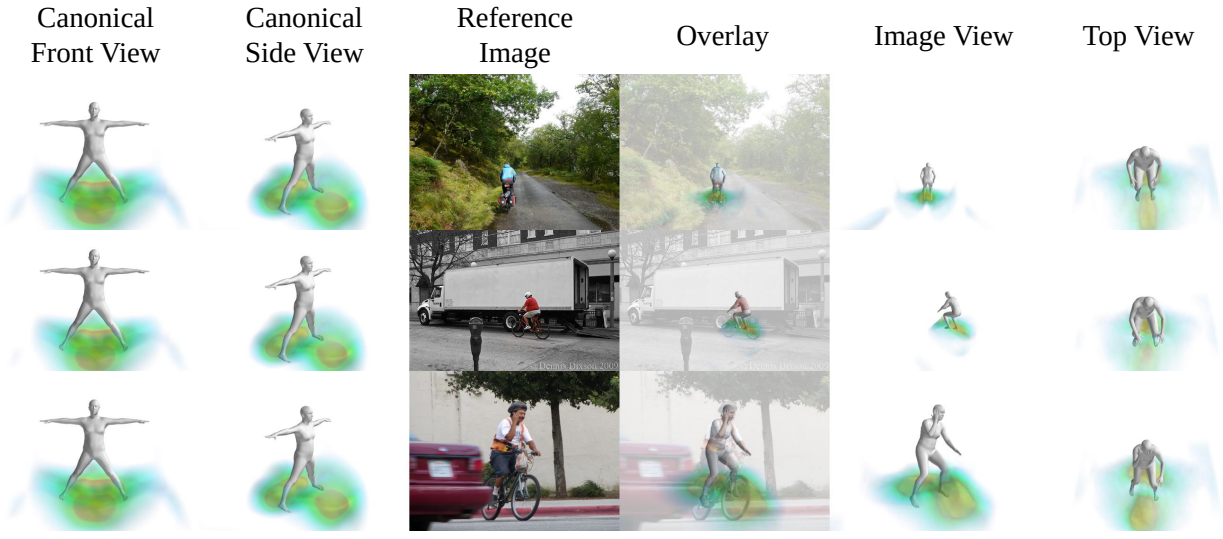

Figure 15. Qualitative results for category **bicycle**.

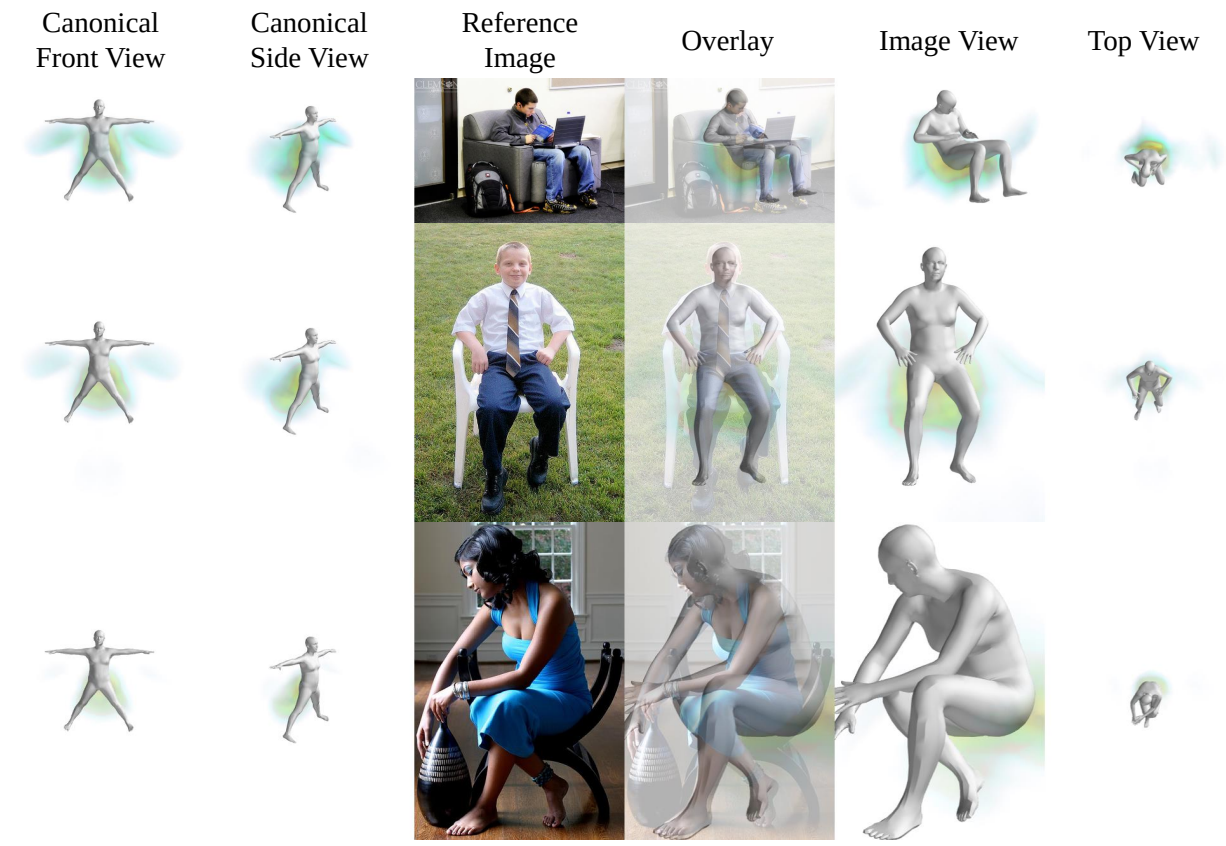

Figure 16. Qualitative results for category **chair**.

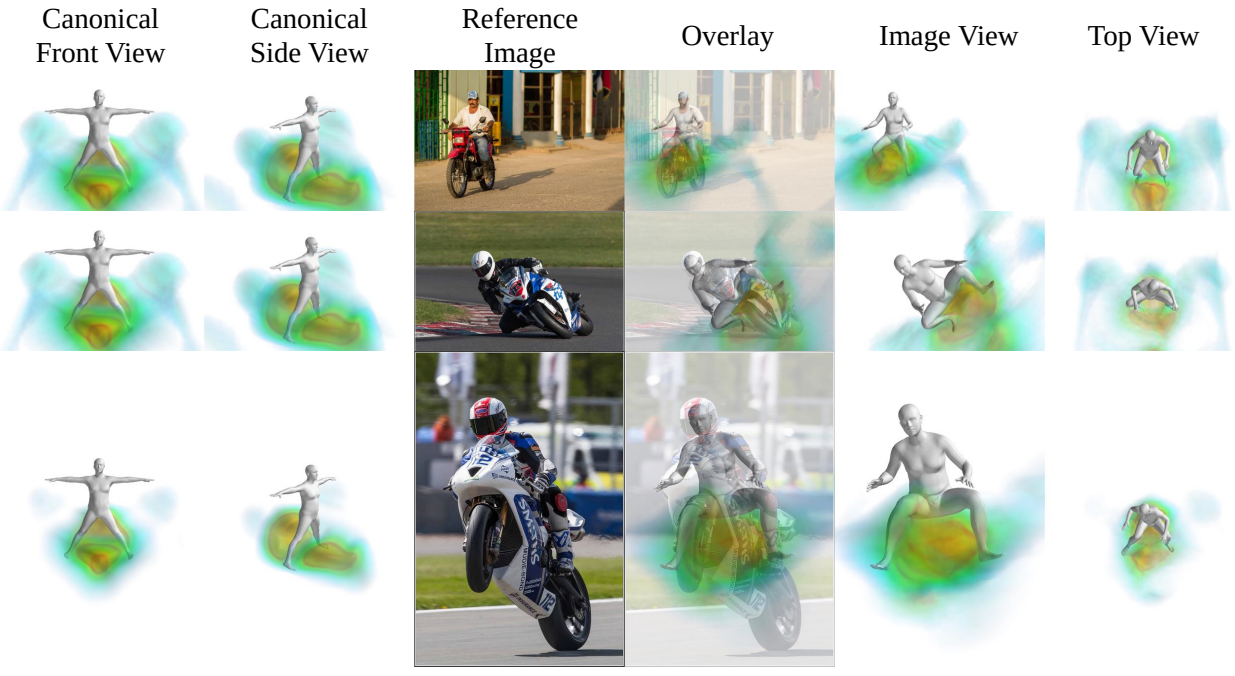

Figure 17. Qualitative results for category **motorcycle**.

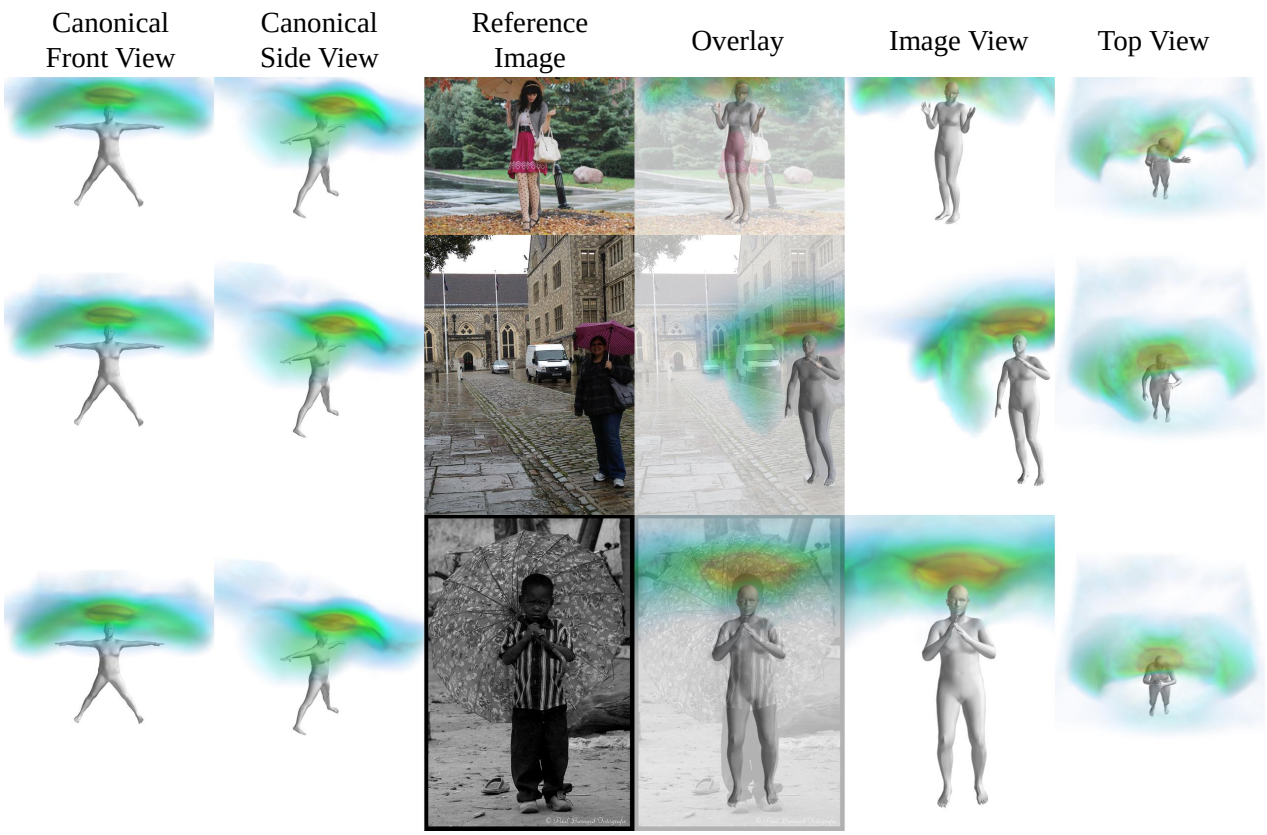

Figure 18. Qualitative results for category **umbrella**.

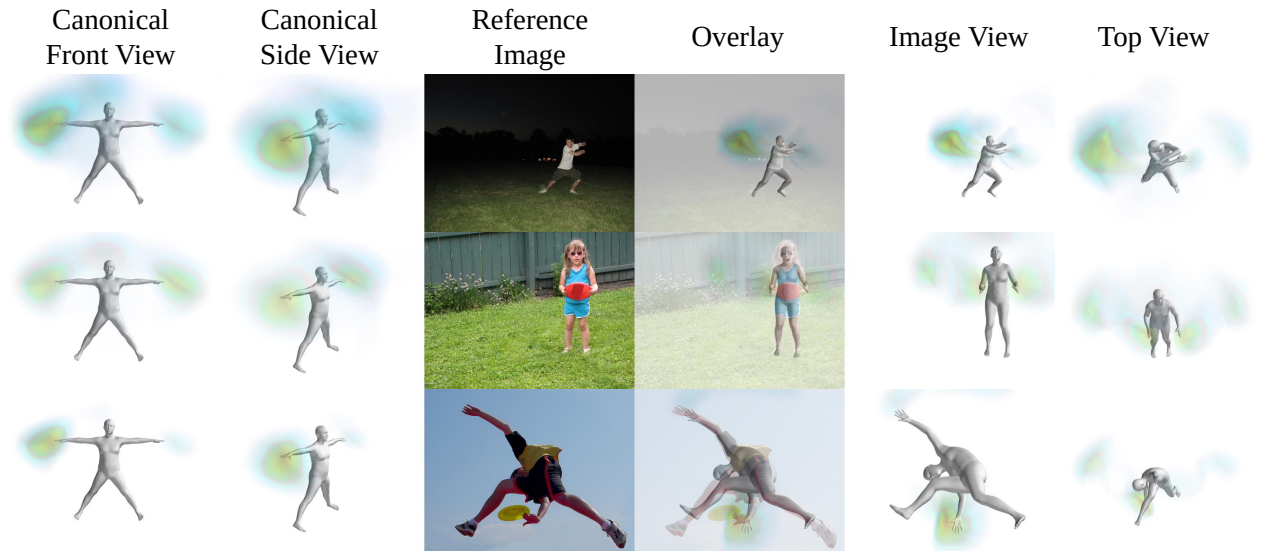

Figure 19. Qualitative results for category **frisbee**.

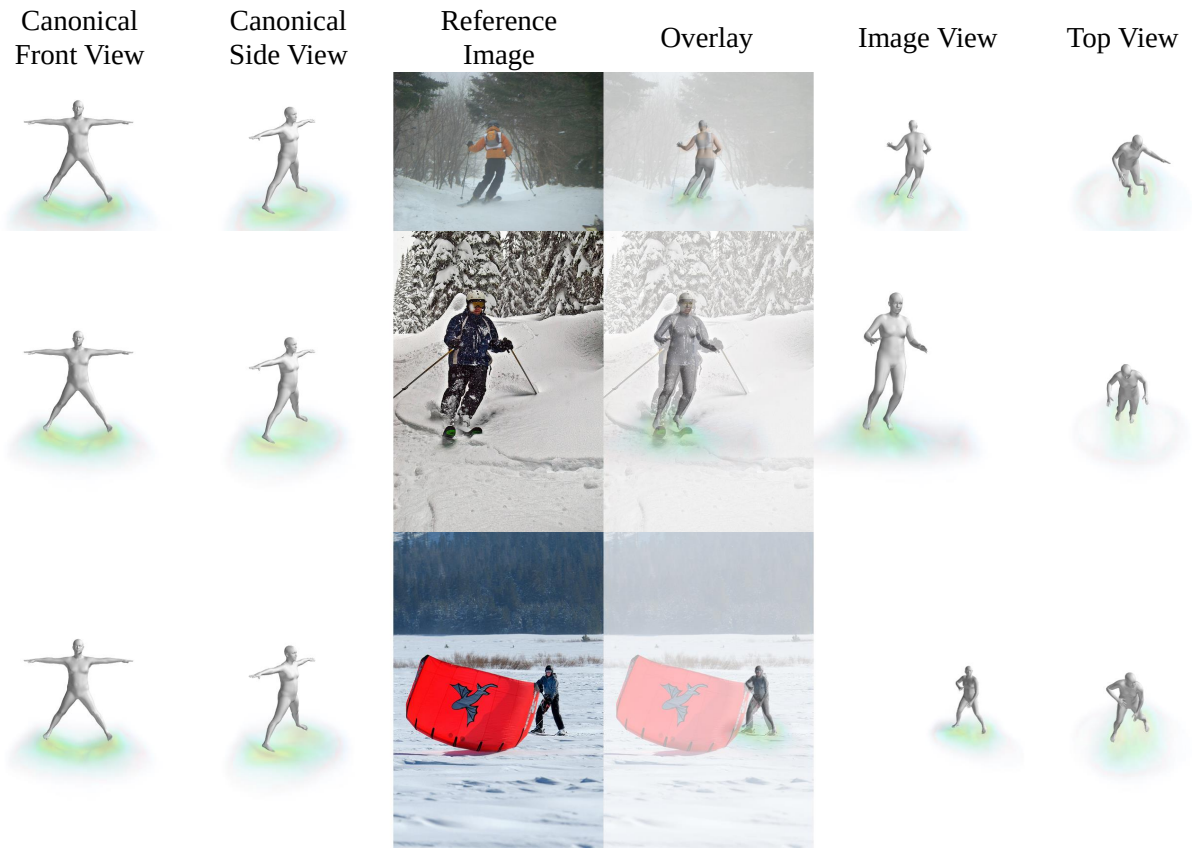

Figure 20. Qualitative results for category **skis**.

Figure 21. Qualitative results for category **surfboard**.

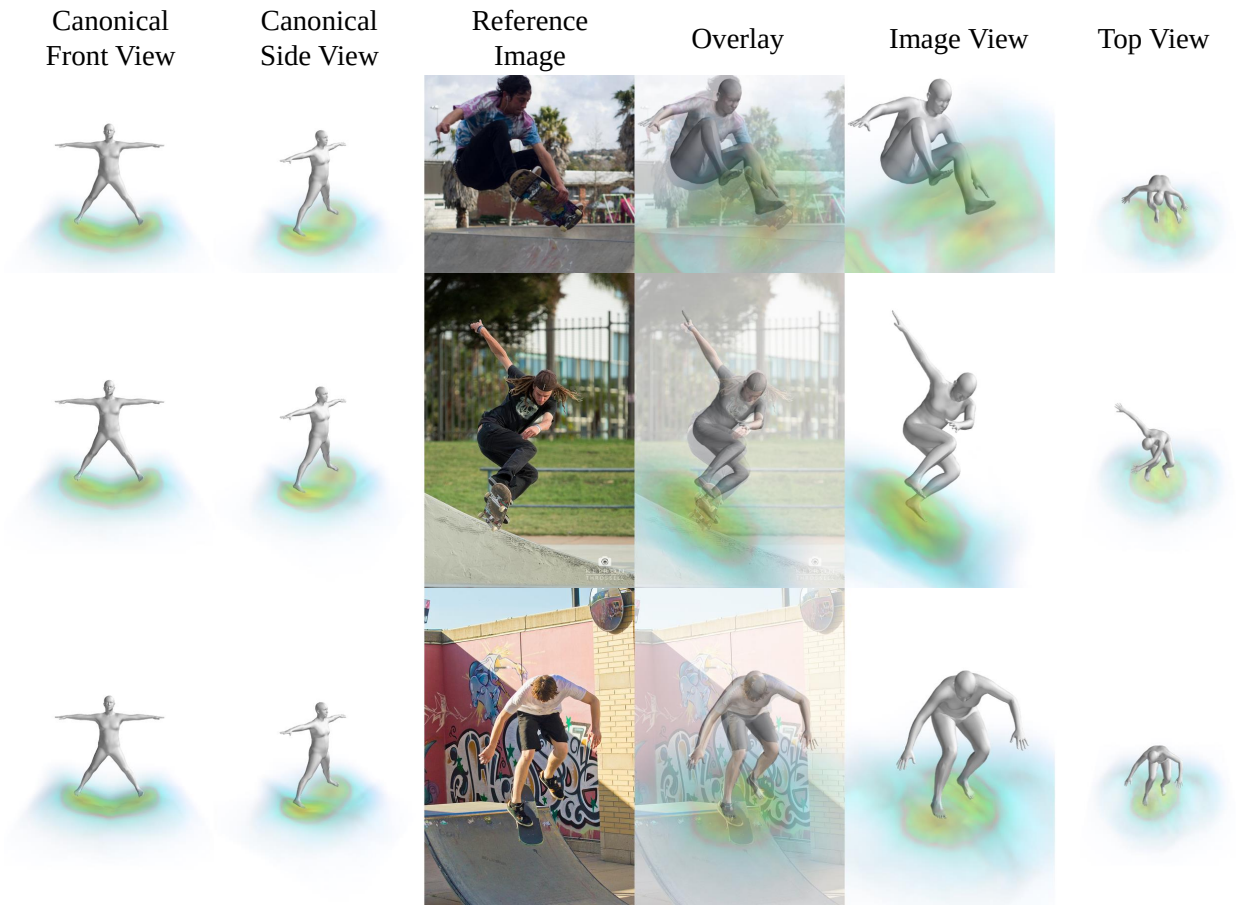

Figure 22. Qualitative results for category **skateboard**.

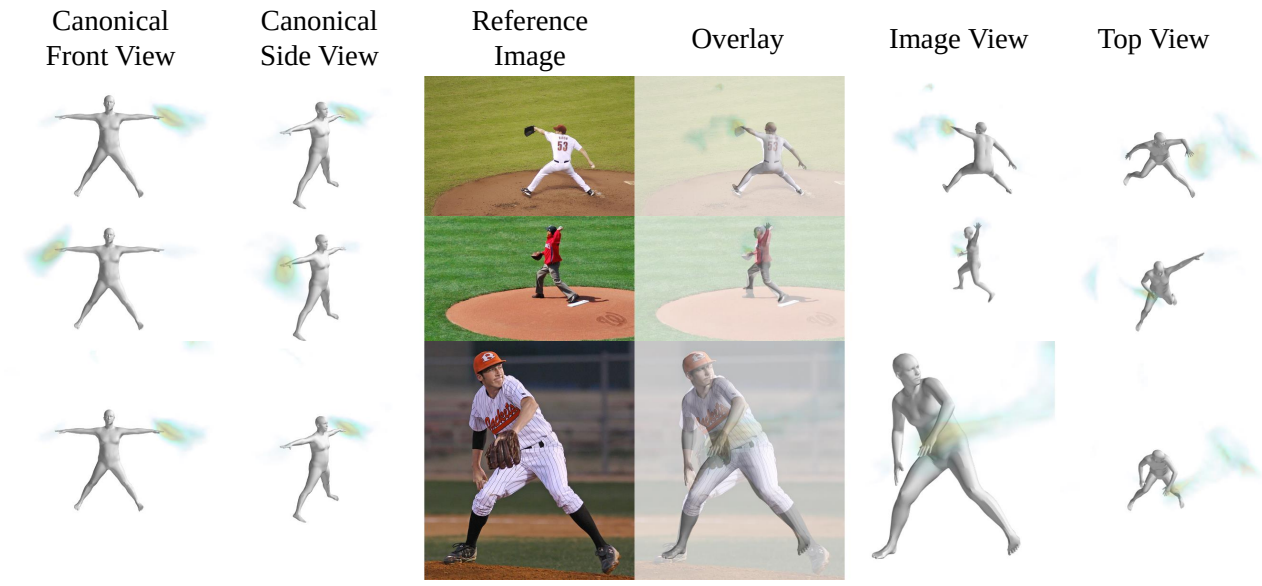

Figure 23. Qualitative results for category **baseball glove**.

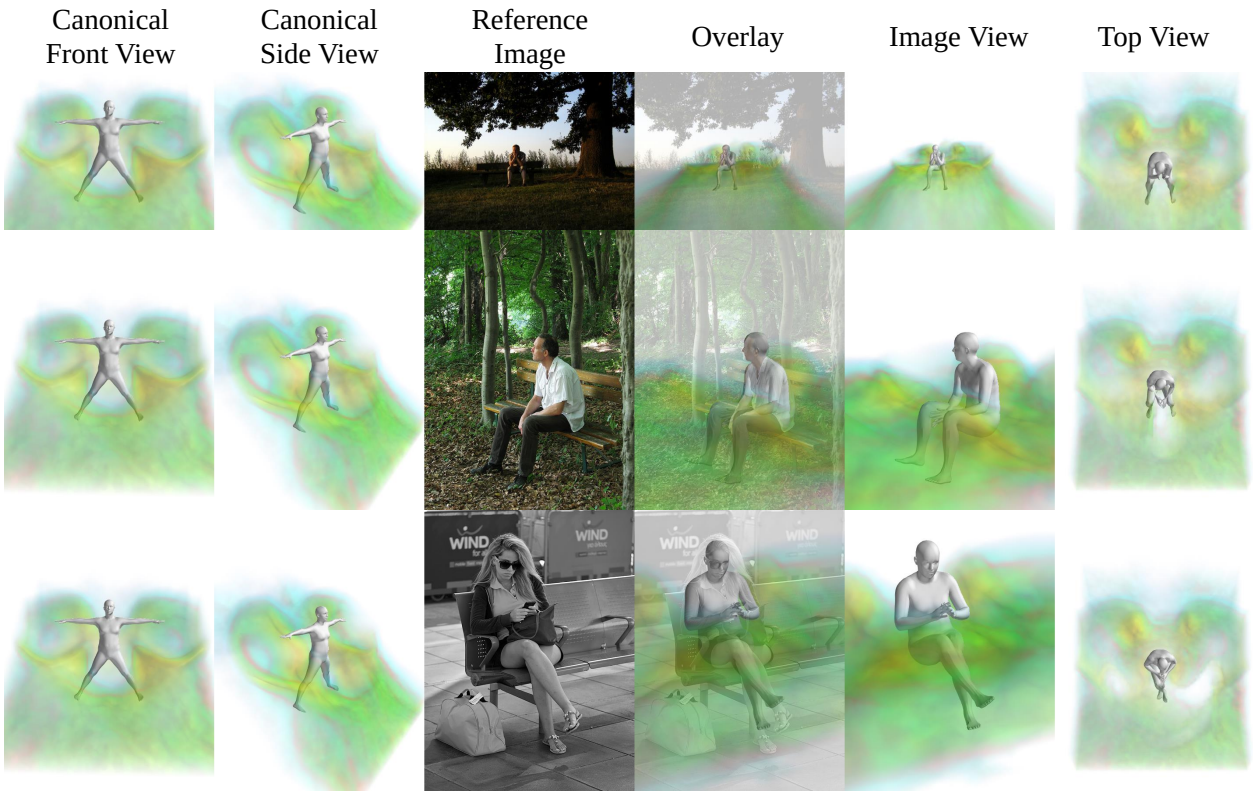

Figure 24. Qualitative results for category **bench**.

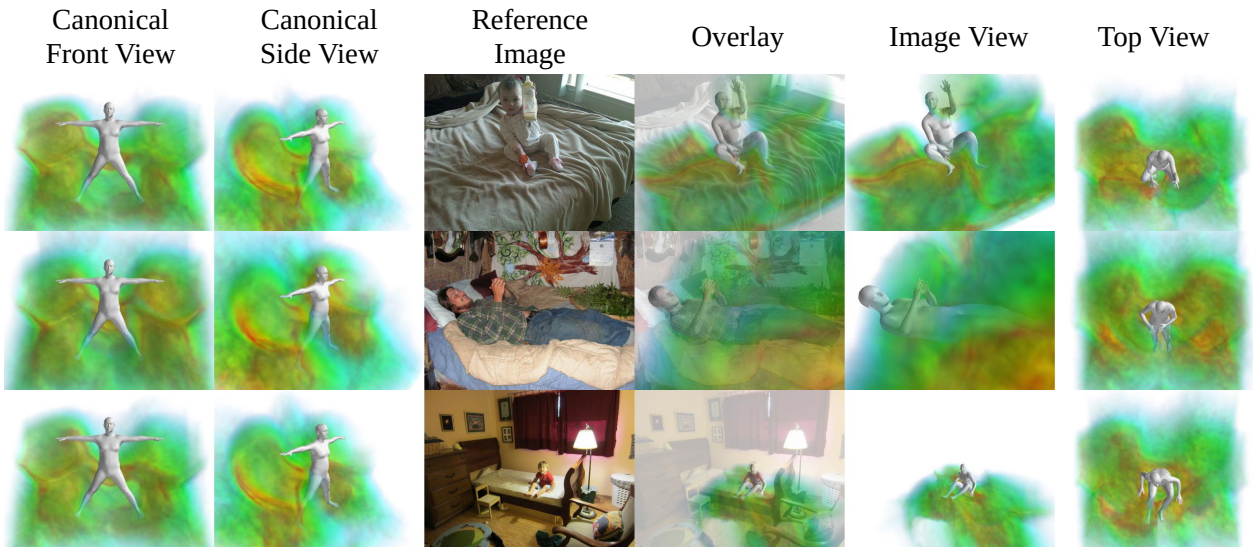

Figure 25. Qualitative results for category **bed**.

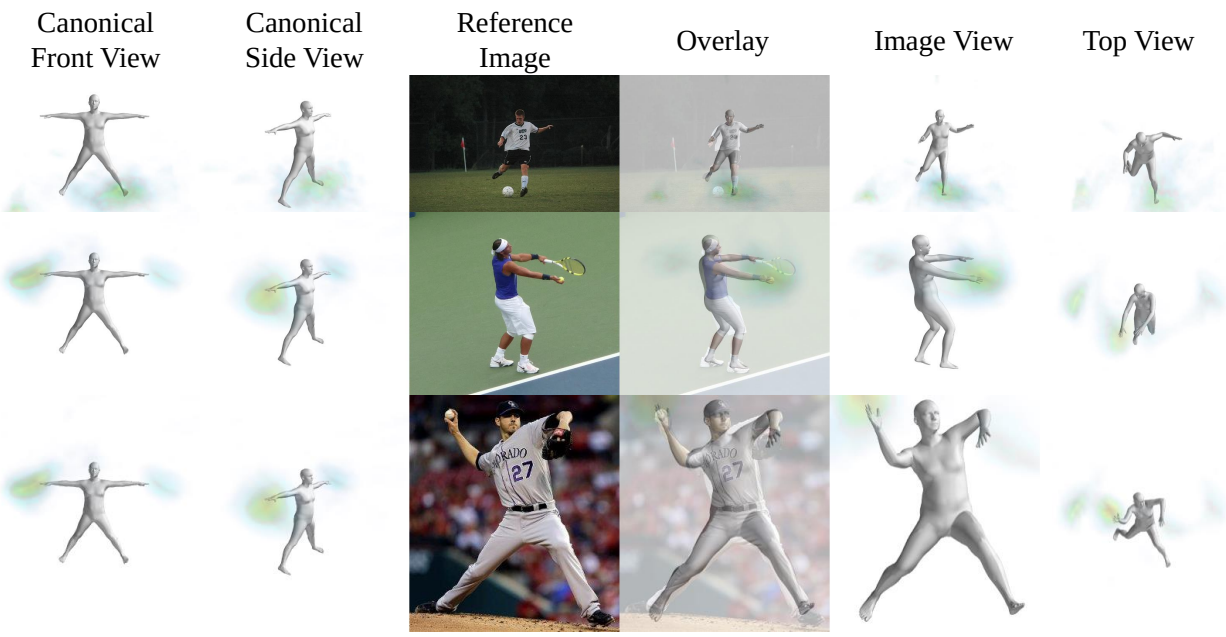

Figure 26. Qualitative results for category **sports ball**.

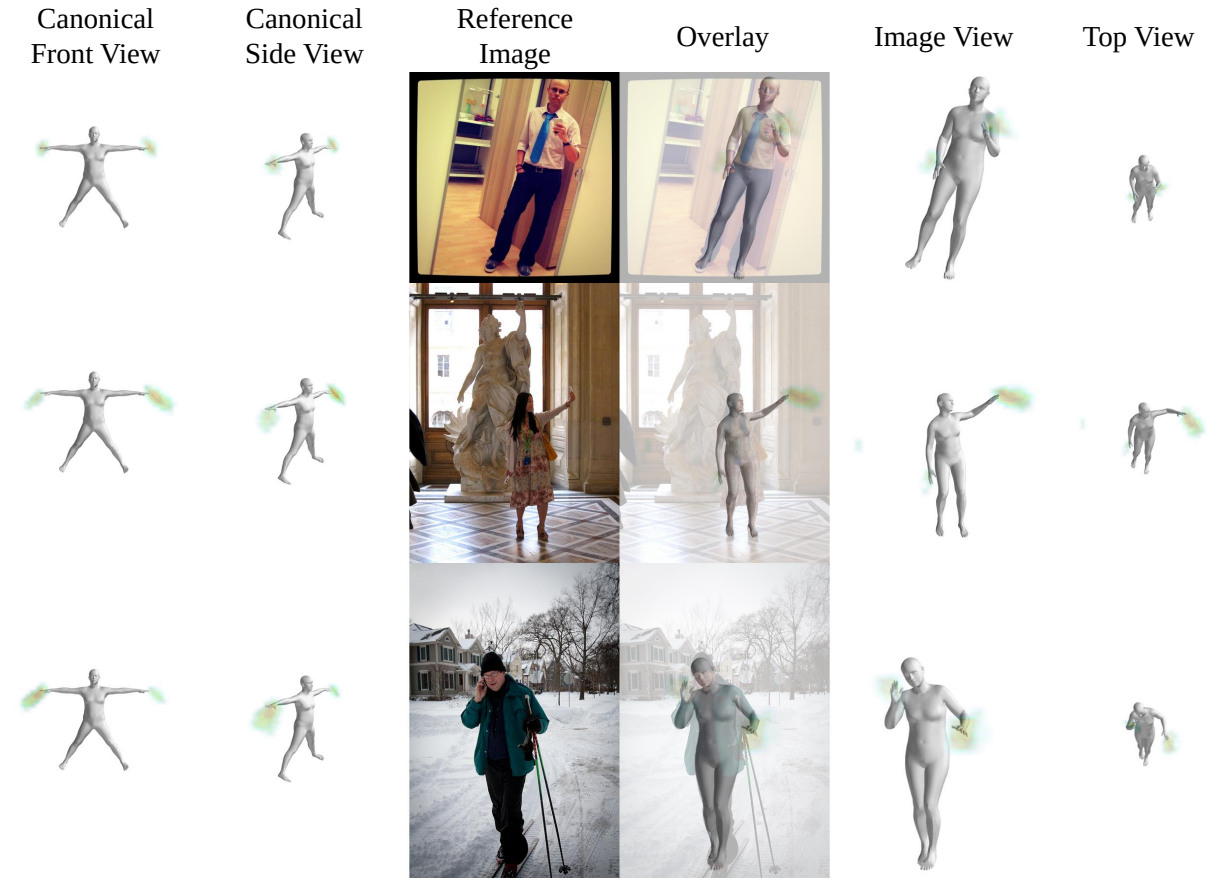

Figure 27. Qualitative results for category **cell phone**.

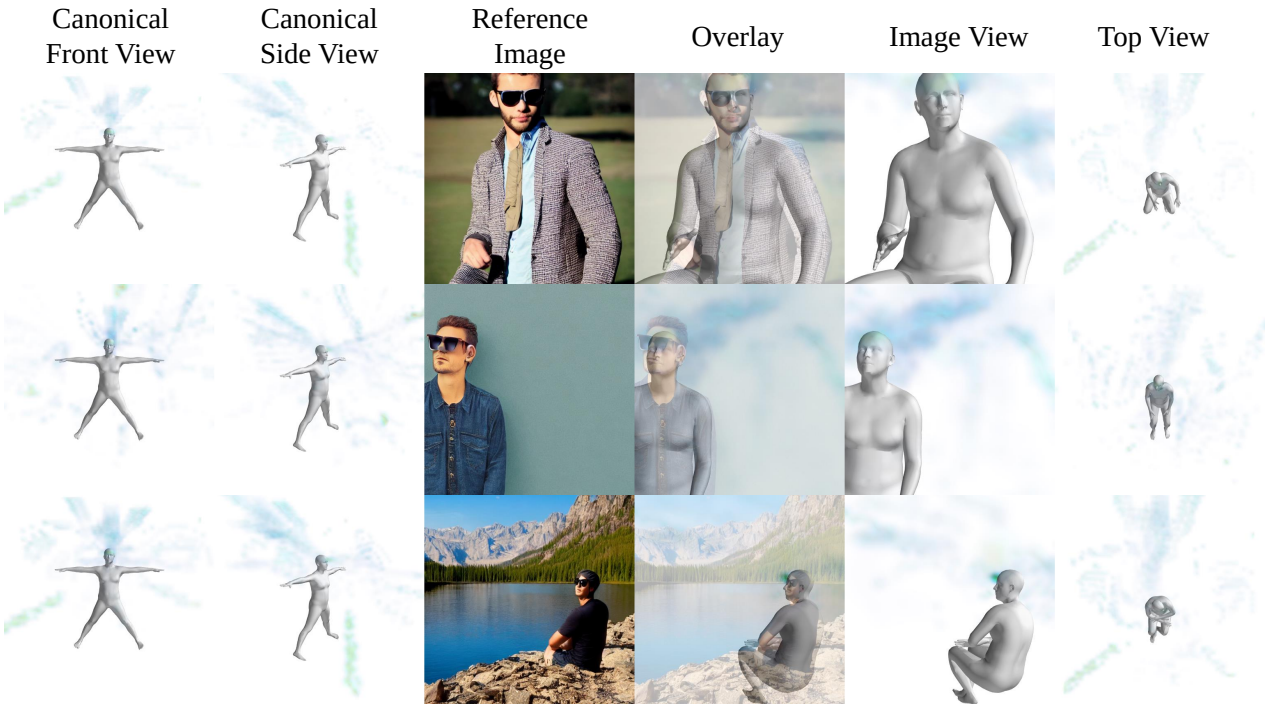

Figure 28. Qualitative results for category **sunglasses**.

Figure 29. Qualitative results for category **soccer ball**.

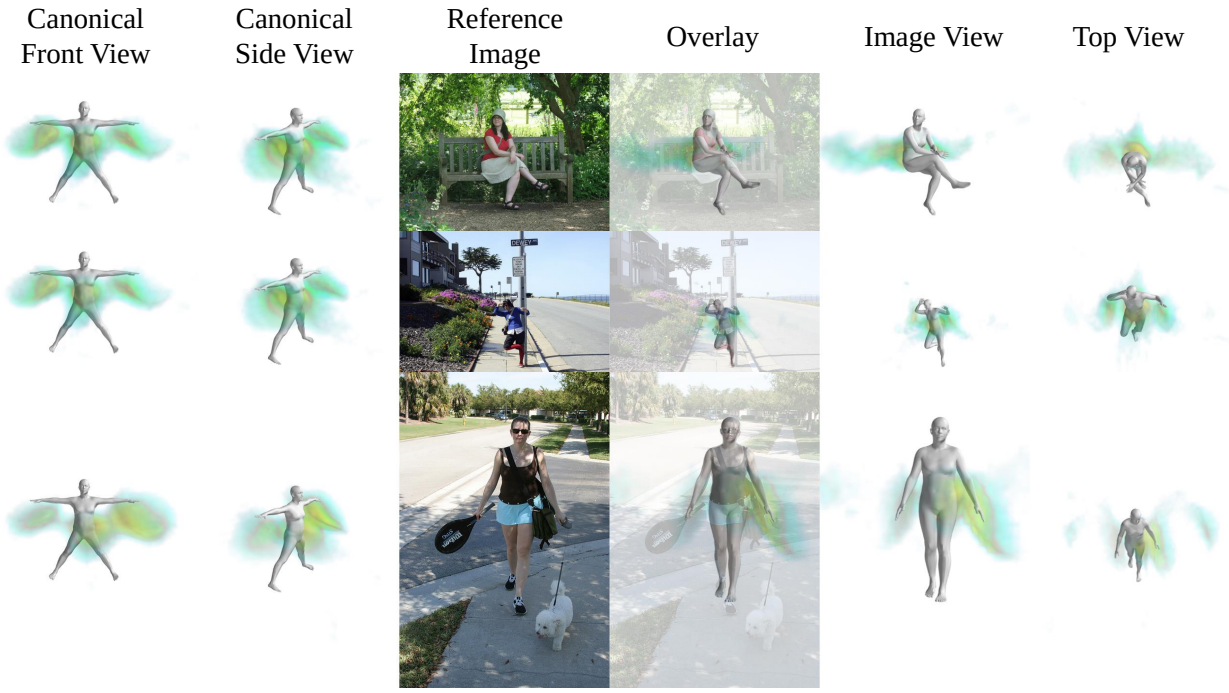

Figure 30. Qualitative results for category **handbag**.

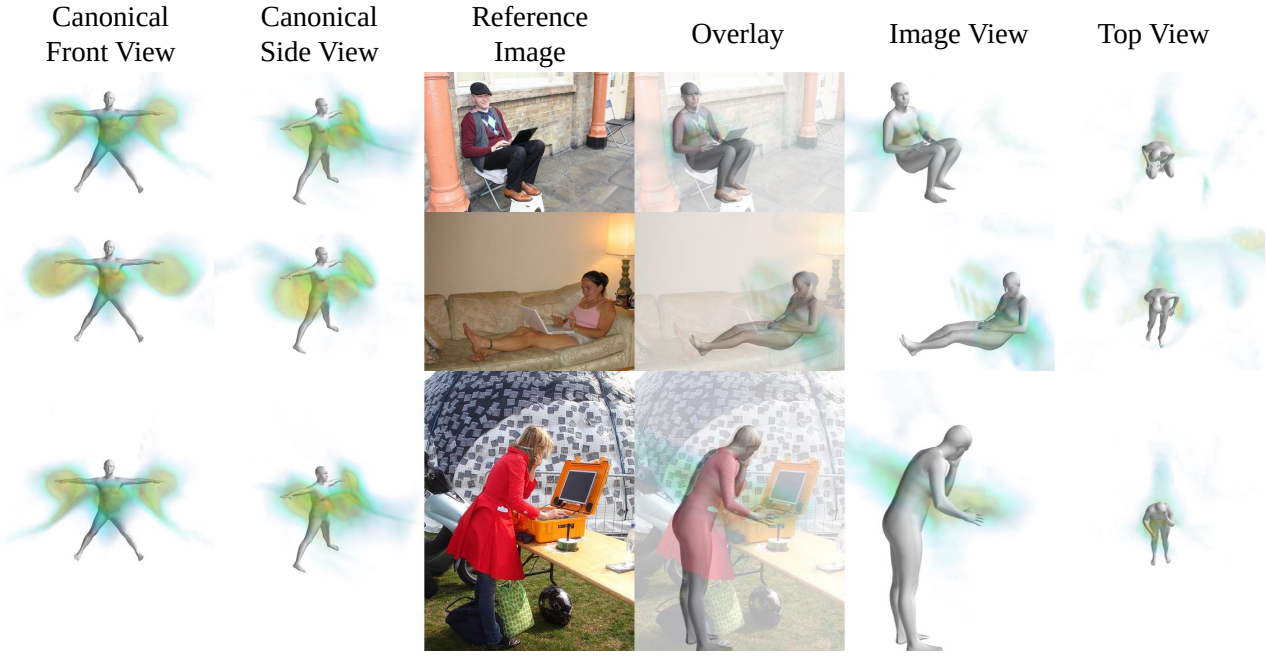

Figure 31. Qualitative results for category **laptop**.

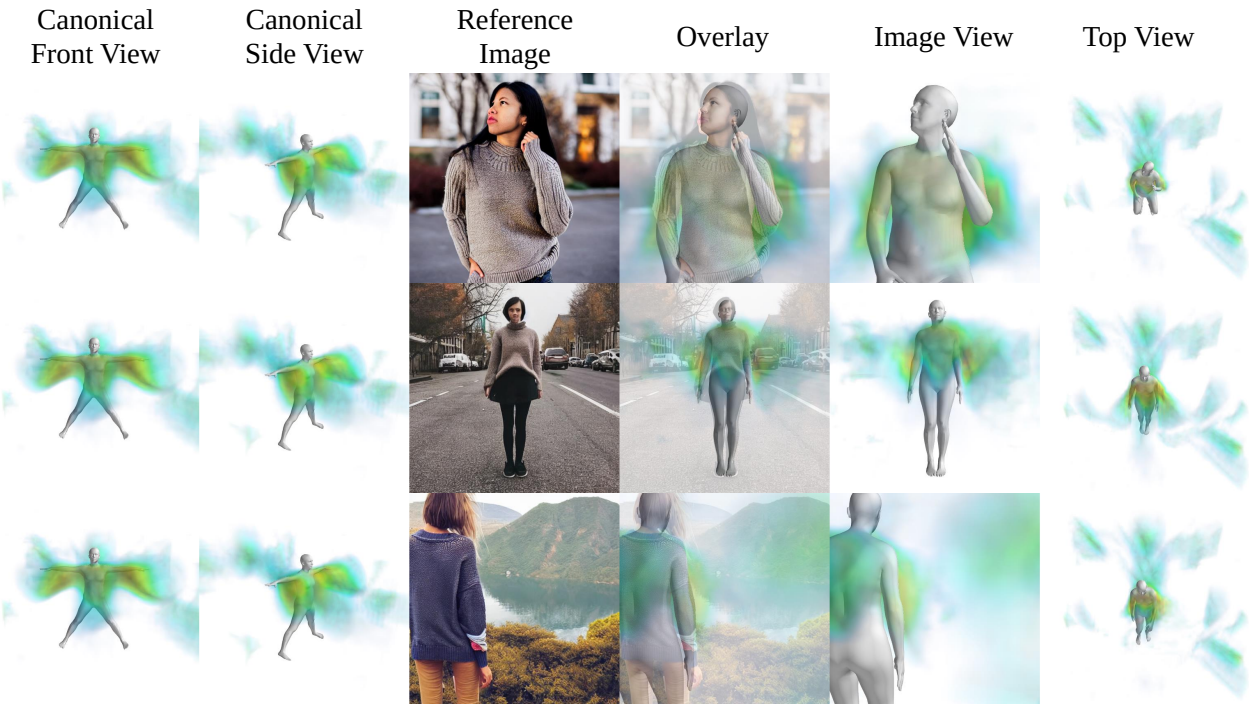

Figure 32. Qualitative results for category **sweater**.

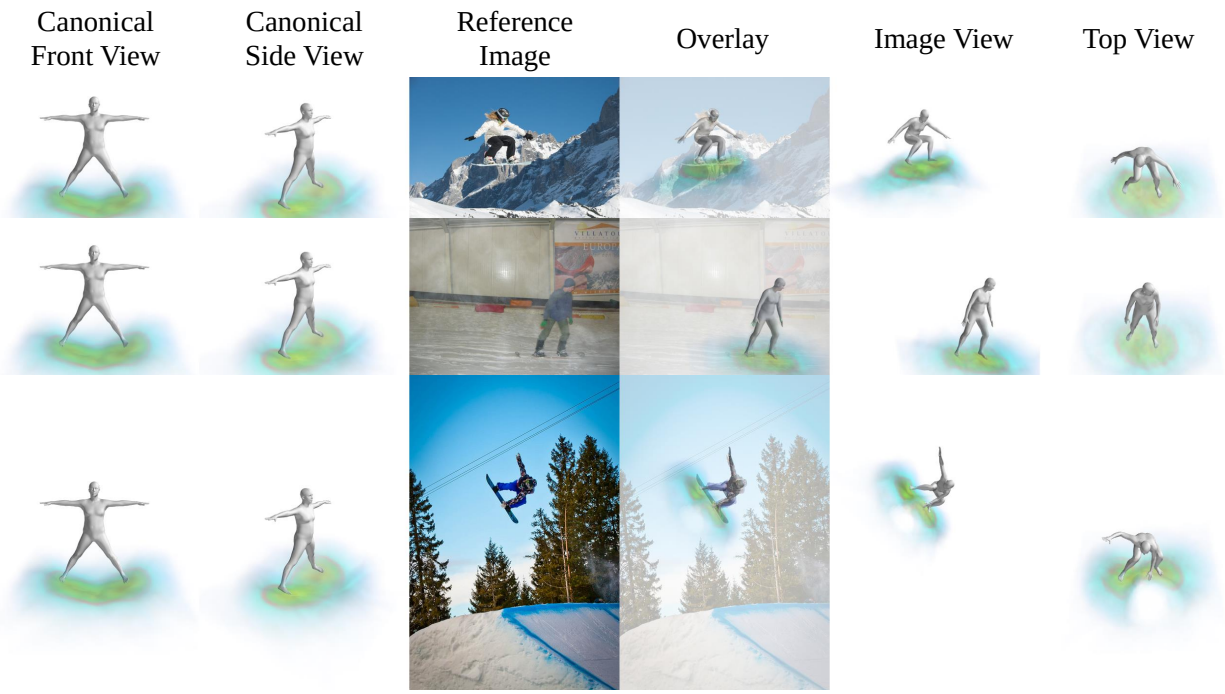

Figure 33. Qualitative results for category **snowboard**.

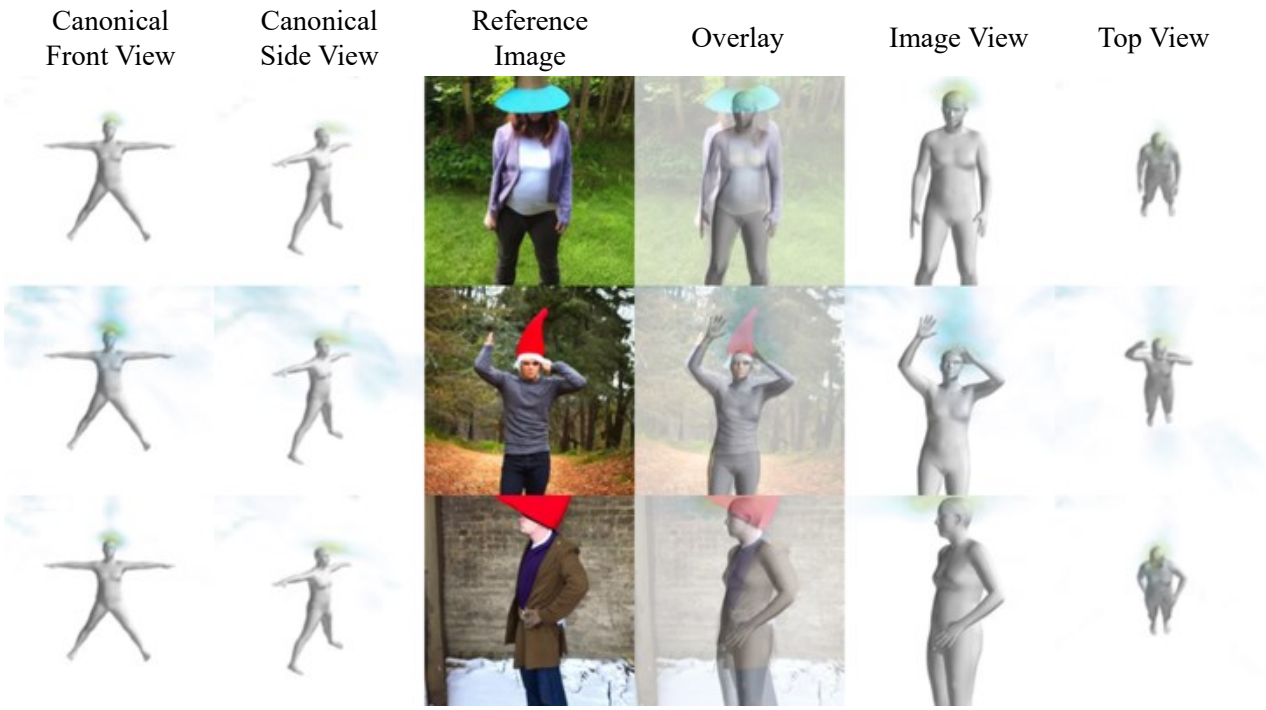

Figure 34. Qualitative results for category **hat**.

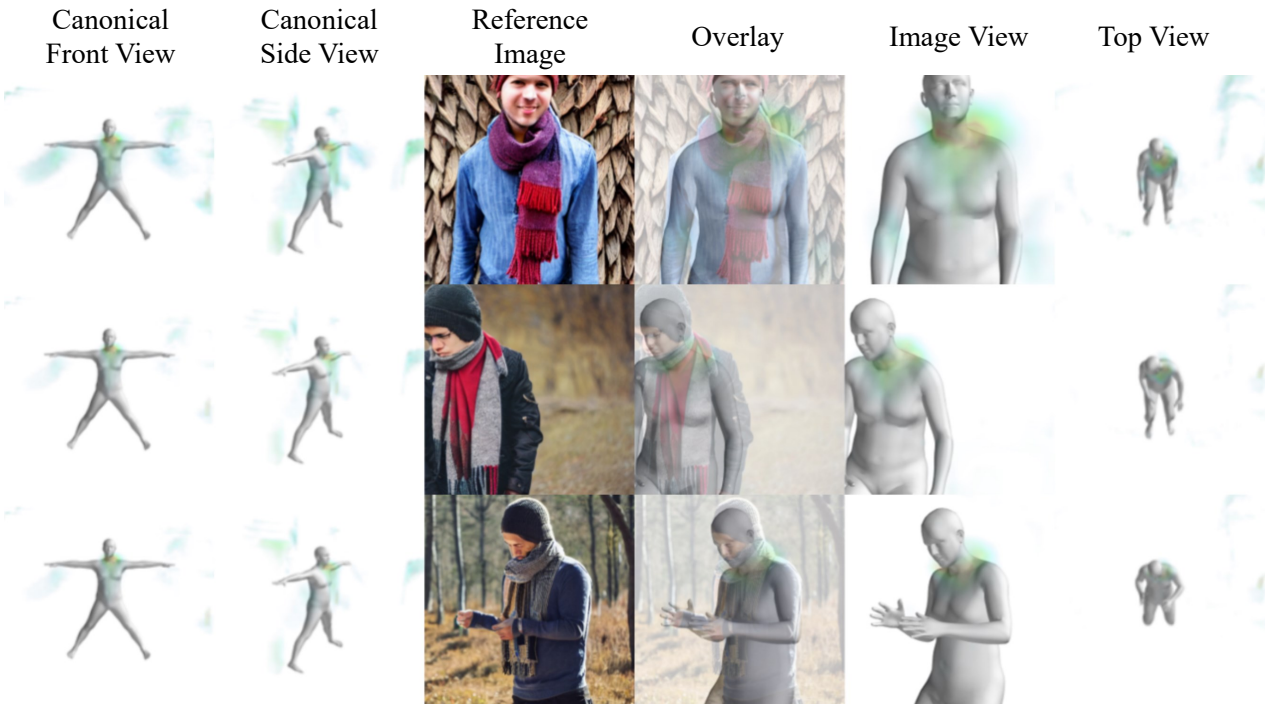

Figure 35. Qualitative results for category **scarf**.

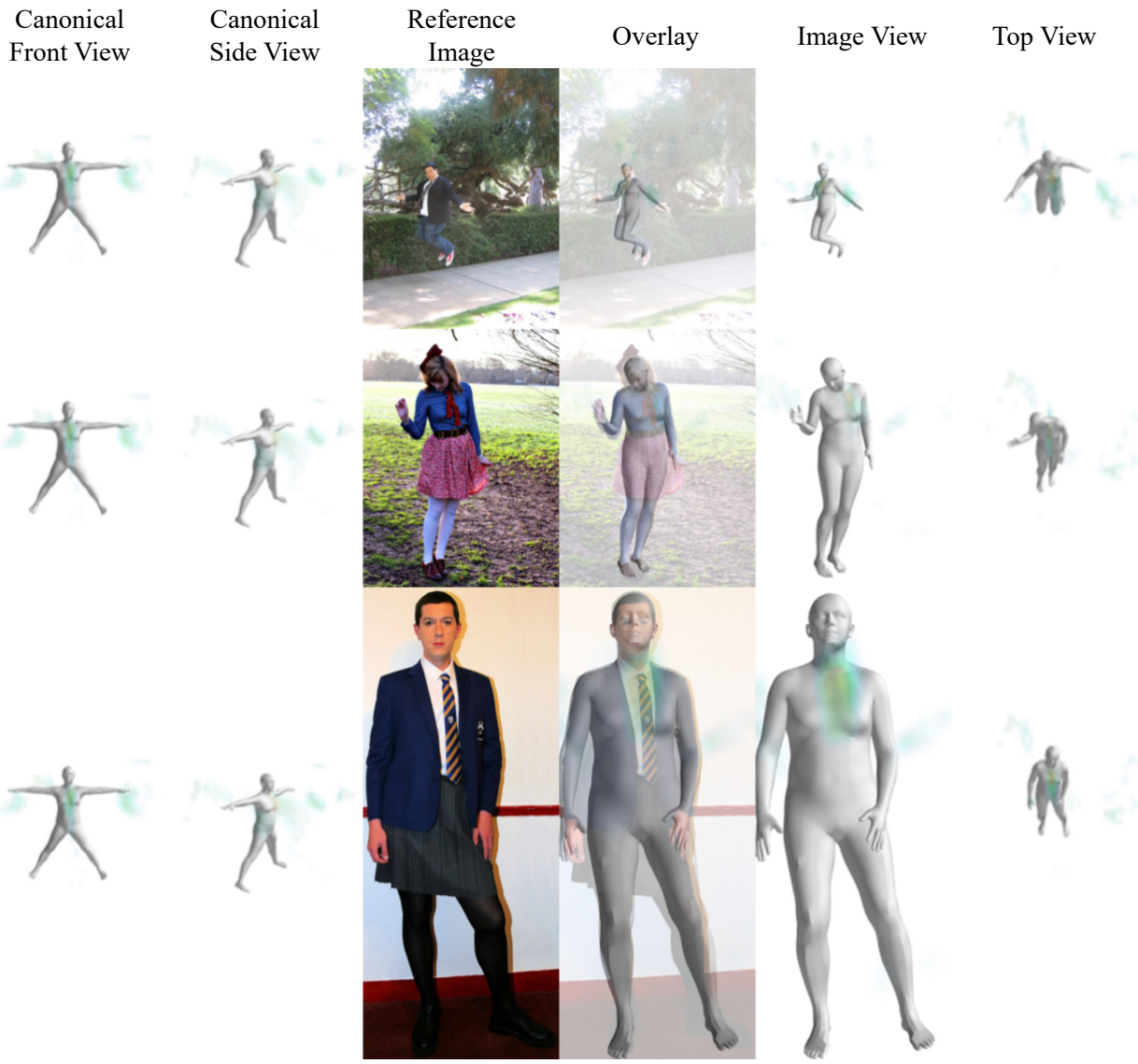

Figure 36. Qualitative results for category **tie**.