IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

VTTB: A Visuo-Tactile Learning Approach for Robot-Assisted Bed Bathing

Yijun Gu and Yiannis Demiris

***Abstract*— Robot-assisted bed bathing holds the potential to enhance the quality of life for older adults and individuals with mobility impairments. Yet, accurately sensing the human body in a contact-rich manipulation task remains challenging. To address this challenge, we propose a multimodal sensing approach that perceives the 3D contour of body parts using the visual modality while capturing local contact details using the tactile modality. We employ a Transformer-based imitation learning model to utilize the multimodal information and learn to focus on crucial visuo-tactile task features for action prediction. We demonstrate our approach using a Baxter robot and a medical manikin to simulate the robot-assisted bed bathing scenario with bedridden individuals. The robot adeptly follows the contours of the manikin's body parts and cleans the surface based on its curve. Experimental results show that our method can adapt to nonlinear surface curves and generalize across multiple surface geometries, and to human subjects. Overall, our research presents a promising approach for robots to accurately sense the human body through multimodal sensing and perform safe interaction during assistive bed bathing.**

I. INTRODUCTION

Robotic bathing assistance could benefit the lives of millions of older adults and individuals with mobility impairment [1], [2]. A robot caregiver offers an opportunity to clean themselves, maintain their privacy, and alleviate the burden on healthcare workers. Yet, robot-assisted bathing presents several challenges, notably in body sensing. The human body contains various dimensions of bones with nonlinear surfaces [3], [4], making it difficult to accurately predict the body shape and model the interaction between the body and the robot. Recent works in robotic bathing assistance have focused on vision-based sensing methods using external cameras [5], [6]. However, due to camera focus limitations, occlusions, and light conditions, visual perceptions could not fully capture the deformation of contact regions, which may cause harmful actions during interactions between robots and humans. Therefore, tactile sensing which captures contact geometry including location and shape, in addition to interaction forces comes as a potential solution to these limitations.

In this paper, we propose a novel method Visuo-Tactile Transformer-based imitation learning for Bathing assistance (VTTB), which the robot learns to track 3D body contours and perceive task-relevant contact representations from expert demonstrations (see Fig. 1). We utilize both depth and tactile images from the demonstration dataset into two cross-attention transformer encoders. The encoder adapts streams from one modality to another, enabling modalities to attend to relevant elements in other modalities and capture long-range cross-modal contingencies. To integrate global contextual reasoning, a standard transformer encoder is applied. We employ a deep imitation learning method that integrates human expertise during learning, aligning the behavior of the robot with human intuition and preferences. We demonstrate our approach with a real robotic system that performs bathing assistance on a medical manikin lying in bed to simulate a robot-assisted bed bathing scenario. We conduct an ablation study to investigate the contribution of each modality and quantitatively compare VTTB against four imitation learning baselines. Furthermore, we showcase how VTTB can generalize to bathe three unseen body parts of the manikin, and to two human subjects.

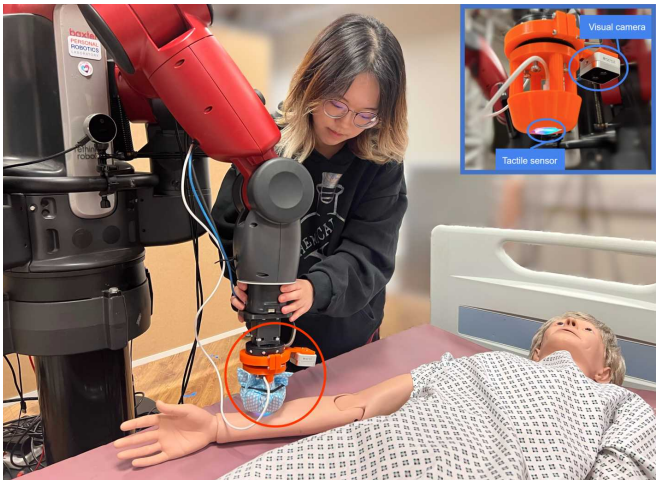

Fig. 1. A human tutor teaches a Baxter robot how to provide bed bathing assistance while simultaneously collecting data from visual sensing, tactile sensing, and robot proprioception. VTTB learns to follow the body contour and perform safe in-contact bathing actions. (Upper right corner) A bespoke cleaning tool which holds a visual camera and a tactile sensor is attached to the robot's end effector.

Through this work, we make the following contributions:

- We present a multimodal sensing approach for assistive bed bathing. We demonstrate how a robot can effectively utilize visual and tactile sensing to interact with a contact-rich bathing environment.
- We propose a Transformer-based imitation learning method that integrates multimodal feedback and learns to focus on important bathing features.
- We evaluate our learned model through multiple exper-

Yijun Gu and Yiannis Demiris are with the Personal Robotics Lab, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom. Emails: e.gu21@imperial.ac.uk; y.demiris@imperial.ac.uk. This work was supported in part by UKRI under Grant EP/V026682/1, and in part by a Royal Academy of Engineering Chair in Emerging Technologies. Videos are available on our project website: `https://www.imperial.ac.uk/personal-robotics/research/vttbathing`.

iments that demonstrate its ability to focus and adapt to contact areas within safe force bounds and generalize across multiple surface geometries. We also present our model bathing two human subjects safely which shows its potential for real-world bathing care.

II. Related Work

A. Robot-Assisted Bed Bathing

Bed bathing is an essential part of nursing care which needs the caregiver's contribution to the comfort, safety, well-being, and dignity of the individual [7]. Human caregivers provide bathing assistance but struggle with significant physical and time burdens due to the growing demand for care. Recent advances have explored real robotic systems to provide physical bathing assistance. King et al. have designed a robotic wiping system that wipes small patches of debris off a person's arm and leg [8]. Erickson et al. present a capacitive sensing approach to track human limb movement and assist cleaning tasks with a wet washcloth [9]. Huang et al. introduce a depth camera-based soft tactile sensor to perform in-contact bathing assistance by leveraging contact surface deformation and user-defined path trajectories [10]. Moreover, Madan et al. present RABBIT, which fuses RGB-thermal perception with compliant control for safe bathing manipulation [11].

Further to these works, our proposed methodology leverages both 3D body contour and task-relevant contact feedback to perform safe contact interaction during assistive bathing.

B. Visuo-Tactile Sensor Fusion

Achieving efficient bed bathing assistance on the human body is challenging since cleaning the body involves direct human skin contact, which needs to be safe and comfortable. Tactile sensing, compared to force sensing, can measure applied force, but also capture contact pressure distribution and spatial information across the sensing surface. Current studies across various in-contact robotic manipulation tasks have researched on integrating vision sensing with tactile sensing to help robot more precisely predict object shapes and grasping forces, leading to improvements in handling [12], grasping [13], [14], in-hand manipulation [15], [16], and insertion [17]. This combined sensory approach also contributes to a more comprehensive understanding of surface properties and textures, thereby increasing accuracy in object recognition [18] and cloth texture identification [19], [20]. Additionally, successful applications in edge or surface following [20], [21] further demonstrate the efficacy of the integration.

In this research, we demonstrate that visuo-tactile sensor fusion can also be used to account for in-contact surface geometry during interactive bed bathing and guide motion prediction development.

C. Imitation Learning for Robotic Manipulation

Conventional approaches for in-contact robotic manipulation tasks largely focus on constraint-based robot programming with control to define the robot's behavior during contact, ensuring stability, safety, and task accomplishment [22], [23]. However, these approaches require application-specific expertise including knowledge on intricate mathematical formulations and precise understanding of robot's dynamics in contact scenarios. In contrast, imitation learning which provides robots an intuitive and easy way to acquire skills directly from human demonstrations, has shown success in these tasks [24]–[26]. Parametric approaches are employed to handle complex input signals, such as vision inputs [27]–[30]. These studies utilize neural networks and build strong assumptions about the parametric state transition distribution or the value function as a function of state-action features. However, common parametric approaches are sensitive to observation noises. Recent advancements leverage Transformer architectures, which prove their efficacy in robot manipulation tasks by integrating an attention mechanism [30]–[34]. Our proposed method extends a cross-attention transformer encoder to a parametric imitation learning policy. This extension improves policy robustness towards visual and tactile observation variations and addresses tasks that involve complicated in-contact interactions.

III. Learning to Bathe

In this section, we present a multimodal learning approach that enables a robot to follow the surface contours of body parts and perform continuous in-contact bathing assistance. We formulate the bathing problem and detail Transformer-based imitation learning approach to address this problem. Lastly, we introduce the observation space, action space, data collection process, and training details used in the bathing task.

A. Problem Statement

We formulate our in-contact bathing task on nonlinear surfaces as an infinite-horizon Markov Decision Process without rewards (MDP\R), defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition dynamics given by $\mathcal{P}(s'|s, a)$. In our robot-assisted bed bathing, $\mathcal{S}$ is the space of the robot's raw sensory data, which includes visual images, tactile images, and proprioception, $\mathcal{A}$ is the action space of the robot's motor commands, $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is a closed-loop sensorimotor policy to perform the task. We assume access to a dataset of N trajectories $D = \{(s_0^i, a_0^i, s_1^i, a_1^i, \ldots, s_{Ti}^i)\}_{i=1}^N$. Our goal is to learn a visuo-tactile servoing control policy π to guide the agent to provide expert bathing assistance, using the demonstrations.

B. Multimodal Transformer-based Imitation Learning

During early pilot studies, we analyze the collected demonstrations provided by a human tutor. As shown in Fig. 3, we observe that the actions depend on the in-contact movement, which contains a temporal relation in sequence. We also notice that the difference between contacts can be small in pixel space, requiring more attention. Based on these initial findings, we employ Transformer as the policy

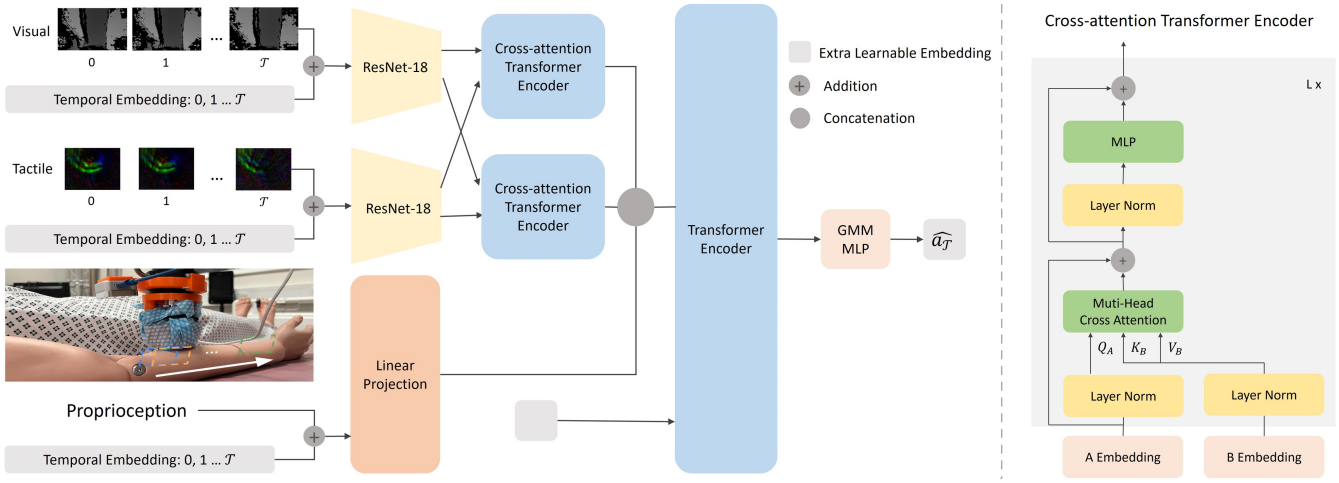

Fig. 2. **An overview of VTTB architecture.** The robot collects visual, tactile, and proprioception observations while sliding over the body surface. Our approach first processes the collected observations into a sequence of feature vectors. Then we feed the visual and tactile data into two cross-attention transformer encoders to enable modality interactions. The two streams of cross-attention are further concatenated with proprioception embedding and an extra learnable embedding to classify action outputs. We proceed all features to a standard transformer encoder to model the global context and then output a latent representation passed through a GMM-MLP to generate actions.

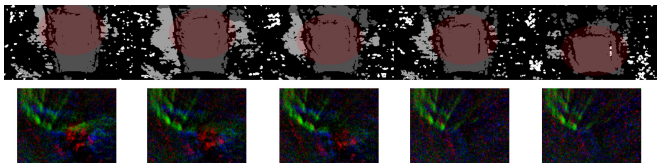

Fig. 3. An illustrative series of visual-tactile observations from the collected demonstrations is presented. From the visual observations, we notice dynamic changes in the position of the red-shaded region-a sample curvature of the body surface. These changes over time sequences reveal a temporal correlation. From the tactile observations, we observe that the differences between contacts are small in pixel space, requiring more attention.

backbone, which has shown the power to draw strong relationships between task-relevant features using a multihead attention mechanism. In this work, we modify a Vision Transformer (ViT) [35] to create a rich temporal distributed latent space. The whole architecture is shown in Fig. 2. We present the details of the architecture in the following sections.

1) Multimodal Observation Embedding: Before being passed into the Transformer, we embed visual, tactile, and proprioception observations into latent vectors. For each visual image $v \in \mathbb{R}^{C\times H\times W}$ where C, H, W are the channel, height, and width respectively, we use ResNet-18 [36] whose convolutional layers spatially combine the features to slice the image into a set of patches. We then feed these patches into a trainable linear projection and add a 1-dimensional positional embedding v^{pos} into $\bar{v} \in \mathbb{R}^{N_v\times D}$ where $N = CHW/P_H P_W$ is the number of the image patches, (P_H, P_W) is the resolution of image patches, and D is the number of embedding features. Similarly, we embed the tactile image t into $\bar{t} \in \mathbb{R}^{N_t\times D}$. For each proprioception, we linearly project it into $\bar{p} \in \mathbb{R}^{1\times D}$.

To learn a temporal relation in sequence, for each step, we encode a sequence $\mathcal{T}$ of past observations as input. Using the above method, we embed visual images, tactile images, and proprioception into $X_V \in \mathbb{R}^{N_V\times D}$, $X_T \in \mathbb{R}^{N_T\times D}$ and $X_P \in \mathbb{R}^{N_P\times D}$ separately, where $N_V = \mathcal{T}N_v$, $N_T = \mathcal{T}N_t$, and $N_P = \mathcal{T}1$. To maintain temporal dependencies of feature vectors, we add temporal positions to each observation group using sinusoidal position encoding introduced by [37].

2) Cross-modal Transformer-based Policy: Given two sequences of input tokens from two modalities X_A and X_B, we modify a standard transformer encoder from ViT to a cross-attention transformer encoder (See right part of Fig. 2) that enables one modality to receive information from another one. The encoder consists of L alternating layers of an Multi-Head Cross Attention (MCA) block, a position-wise MLP block, and Layer Normalization (LN):

$$
\begin{aligned}
Z_A^0 &= X_A; Z_B^0 = X_B \\
Q_A &= \mathrm{LN}(Z_A^{l-1}) && l = 1\cdots L \\
K_B, V_B &= \mathrm{LN}(Z_B^0) \\
\hat{Z}_A^l &= \mathrm{MCA}(Q_A, K_B, V_B) + Z_A^{l-1} && l = 1\cdots L \\
Z_A^l &= \mathrm{MLP}(\mathrm{LN}(\hat{Z}_A^l)) + \hat{Z}_A^l && l = 1\cdots L
\end{aligned}
$$

An MCA block consists of a sequence of cross-attention layers that model pairwise relations between tokens. We first linearly map input tokens into query (Q_A), key (K_B) and value (V_B) representations. Through (1), the layer then computes a weighted sum over the token values, and the weight assigned to V_B is a compatibility function of Q_A with the corresponding K_B. The dot-product is then scaled by $\sqrt{d_K}$ and applied to a softmax function to obtain the attention given by modality A to modality B. To allow the model to attend to information from different combinations of input space representations, the block runs a sequence of cross-attention layers in parallel and projects their concatenated outputs h times with a different set of weights. The outputs from the MCA block are then fed into an MLP block. LN is applied before both MCA and MLP blocks and residual connections after every block.

$$\text{Attention}(Q_A, K_B, V_B) = \text{softmax}(\frac{Q_A K_B^T}{\sqrt{d_k}})V_B \quad (1)$$

In our work, we pass both vision information X_V and tactile information X_T to each other through two cross-attention transformer encoders. The outputs from the encoders are concatenated with the proprioception input tokens X_P. We further append a learnable embedding of an aggregation token whose corresponding output is used as an aggregated representation for the entire input sequence. We use it for action prediction. To further model the global context, we process the combined vector into a standard transformer encoder. The token output from the encoder is then passed through a GMM-MLP prediction head [38]. The prediction head is composed of distinct components, beginning with an MLP with two hidden layers, each containing 1024 neurons. Linear neural networks are used to map the output from the MLP to parameters of Gaussian distribution including means, covariances, and logits. Subsequently, a Gaussian Mixture Model (GMM), known for its efficacy in extracting multimodal mixed-quality information from demonstration data, is employed to sample policy action predictions. The resulting predictions are passed through a tanh layer for normalization to [-1, 1].

C. Training the Model

1) Observation Space: The robot receives three kinds of observations: (1) proprioception observations (7-dim) consisting of the end effector position (3-dim), and quaternion (4-dim) relative to the starting end effector pose; (2) depth image observations of the front body surface with a resolution of 212×120. The raw camera frames are read at a full resolution of 848×480; (3) tactile image observations of the contact region between the cleaning tool and the body surface with a resolution of 160×120. The raw camera frames are read at a full resolution of 320×240. Our tactile image observation is processed by subtracting the initial raw camera frame from the current raw camera frame. Both image observations are applied a pixel normalization to a range of (0, 255) and resized to the appropriate resolution.

2) Action Space: We define an action space for the bathing manipulation task in which the robot performs in-contact bathing actions along the surface of body parts following the body contour. Action is designed as a 6-dimensional vector where the first three coordinates are the desired delta translation from the current end effector position, and the other three coordinates are the desired delta rotation in an axis-angle form from the current end-effector rotation. The learned policy outputs actions, which are then transformed into end-effector target poses and sent to the robot to achieve the desired Cartesian poses. With a position-based controller, the learned policy effectively controls the end effector, enabling accurate tracking of the surface contour via visual feedback and optimization of pressure applied to the body surface via tactile feedback. The actions are zero-centered and scaled within the range of $[-1, 1]$ across the training dataset.

3) Data Collection: We collect our demonstration data a total of 144 times on four body parts: upper arm, lower arm, upper leg, and lower leg. Twelve people (6 female, aged 20-31 years) have participated in the study.

For each trial, at the beginning, the manikin is positioned in a supine pose with randomized joint variations. Joint poses (e.g., shoulder, elbow, thigh, or knee) are estimated through a head-mounted camera using the High-Resolution Net (HRNet) library [39]. A start pose is randomly sampled from a uniform distribution from the joint pose. The robot then moves its end effector to this start pose. Subsequently, human tutors manually guide the robot to bathe through the surface of the body part, under the robot built-in kinematics teaching mode. For every 0.1 seconds, the visual data, tactile data, and robot proprioception are recorded simultaneously. We employ a force/torque sensor (Robotiq FT300) to ensure that the human tutors provide demonstrations within a safe force boundary of 10N. Note that the force/torque sensor is only used for data collection and model evaluation. Our goal is to learn a controller which provides safe in-contact bathing actions directly from vision and tactile information.

Data augmentation For each demonstration trajectory, we create five subsets to augment the data. To encourage larger robot movement per time step, we make the subset by picking each frame in every five consecutive frames. To increase the generalization ability of the model, we apply image transformations and pixel shifting [38], which crops the source image at random locations. We also flip and rotate the image to introduce more diversity in the data.

4) Training Details: In this study, we employ Behavioral Cloning (BC) [40] as our imitation learning algorithm. Considering GPU memory constraints, we split both visual and tactile images into 5x6 patches. We choose h=6 parallel attention heads to spread attention distribution and L=6 Transformer layers. The size of the input embedding feature vector D is set to 384. The number of GMM modes is set to 5 with a standard deviation minimum clipping of 1e-4.

All the policies are trained for 200 epochs with a batch size of 16. We follow the general hyperparameters provided by Mandlekar et al. [38] and fine-tune the learning rate, actor network dimension, and time sequence length $\mathcal{T}$ to 1e-5, [1024, 1024] for MLP, and 10 respectively. To train with a GMM-MLP prediction head, negative log-likelihood is used as the loss function. The overall loss term for a given training sequence is the loss averaged across all $\mathcal{T}$ predictions. We utilize the Adam Optimizer [41] and save the checkpoint with the lowest validation loss across the demonstration dataset.

IV. Experiments and Results

Our experiments aim to examine the effectiveness of our proposed method in assistive bed bathing where the robot wipes through the body part's surface with a cleaning cloth. The task requires the robot to maintain continuous contact with the body surface while following the surface contour. We conduct three experiments to investigate: (1) Does visuo-tactile sensing help the robot better perceive the bathing environment? (2) How well does Transformer-based learning

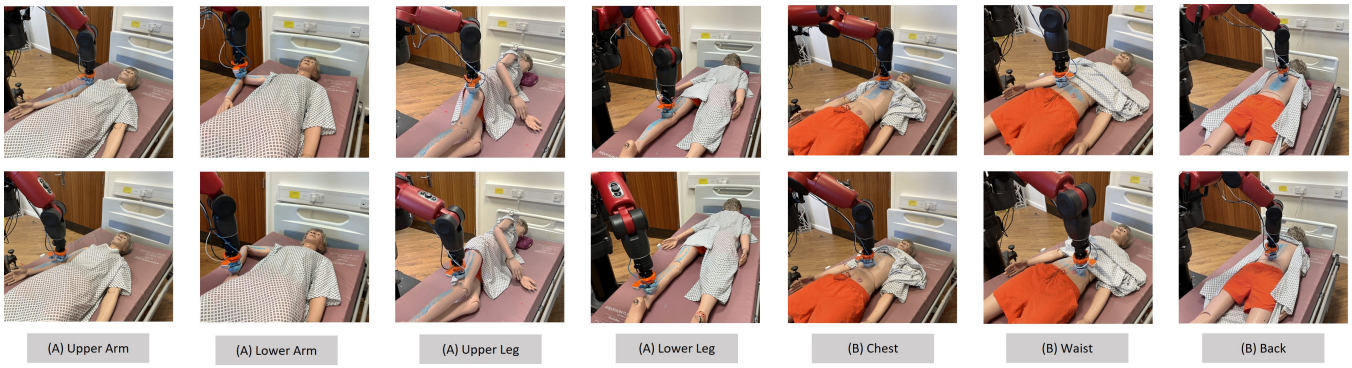

Fig. 4. Snapshots of Baxter robot providing bathing assistance for different body parts in different poses for (A) Multimodal Ablation Study and Baselines Comparison (B) Ability of Generalization. The example trials are also provided on our project website.

perform against other baselines in this task? (3) Does the learned policy generalize over body parts variations?

A. Experimental Setup

We collect data and evaluate our proposed approach using a real Baxter humanoid robot on a bed bathing task (See Fig. 1). We use a professional training medical manikin lying on a hospital bed to simulate an adult person. We design and 3D print a bespoke cleaning tool attached to the robot's end effector for assistive bed bathing. The tool holds a visual sensor and a tactile sensor with a fixed offset between each other. We wrap a thin layer of cleaning cloth over the tactile sensor to simulate the bathing task. Blue powder is put on the surface of the body parts to simulate the dirt to be bathed off. In this paper, we use an Intel Realsense D405 camera to capture depth images of the surface to be bathed and a DIGIT tactile sensor [42] to collect dense visual information regarding the contact region between the tool and the body surface. The sensors are operated at a rate of 30Hz. The learned policies provide predictions for the given visuo-tactile information at an average frequency of 80Hz, enabling the generation of high-level control input at a corresponding frequency to that of the received images.

B. Evaluation Metrics

We evaluate the robot's ability to follow the contours of the body parts and perform continuous in-contact bathing movements. We test the learned policies on four body parts: upper arm, lower arm, upper leg, and lower leg. For each body part, we repeat the experiment 20 times with randomized body poses and start poses. Unlike the data collection process where we only consider the manikin in a supine bed pose, we introduce broader pose variations, including supine, prone, supine or prone with arm and leg flexion, and lateral recumbent, to evaluate the robot's robustness in bathing various body surfaces. Fig. 4 portrays a sequence of snapshots displaying successful bathing trials for each body part in different poses.

We report the quantitative performance of the policies using the task success rate across the body parts. We define the success rate into three different phases:

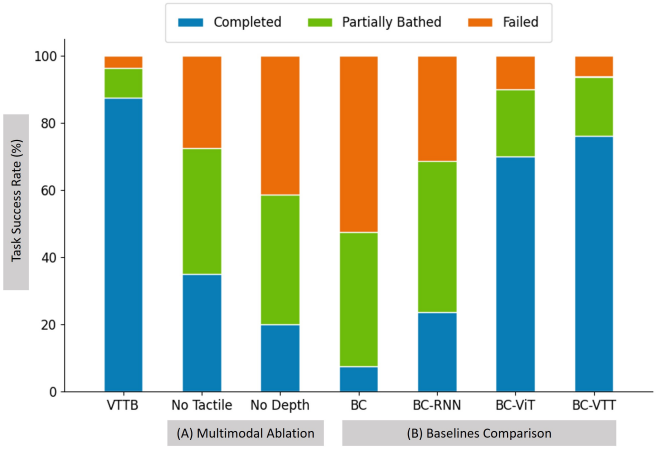

Fig. 5. (A) We conduct an ablative study of policies trained on different combinations of sensory modalities. We compare *VTTB* with *No Tactile* that is trained without tactile images and *No Depth* that is trained without depth images. The graph shows that policy performance drops due to the lack of tactile images and depth images. (B) We compare *VTTB* against four imitation learning baselines: *BC*, *BC-RNN*, *BC-ViT*, and *BC-VTT*. The graph shows that *VTTB* outperforms all baselines, which validates its efficacy.

- *Completed*: The robot bathes along the surface of the body parts and maintains continuous contact with the body.
- *Partially Bathed*: The robot bathes part of the body and moves away from the body.
- *Failed*: The robot fails to bathe the body.

C. Multimodal Ablation Study

In our approach, we utilize depth images and tactile images in addition to robot proprioception. To investigate the contributions of each modality for the bathing task, we conduct an ablation study where we train our policies with three combinations of modalities: (1) *VTTB*: our proposed model; all the modalities are fed into the network (2) *No Tactile*: the tactile image inputs are masked out; only depth images and proprioception are fed into the network (3) *No Depth*: the depth image inputs are masked out; only tactile images and proprioception are fed into the network. For both *No Tactile* and *No Depth*, our cross-attention transformer encoder is the same as a standard transformer encoder.

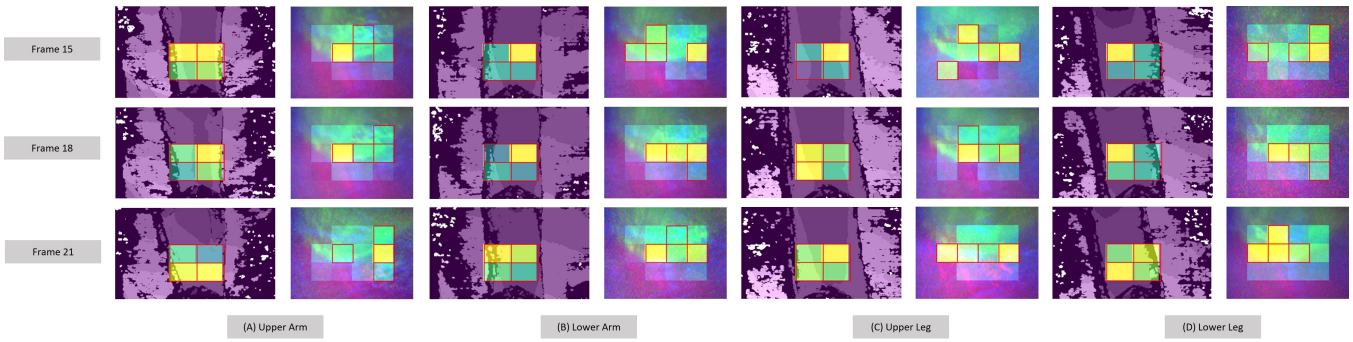

Fig. 6. **Attention Maps** We visualize the temporal attention at the 15th, 18th, and 21th time frames. For each body part, we overlay the attention heatmaps onto both depth images (left) and raw tactile images (right). The top-4 attention tokens of each image are highlighted with red lines. From the visual attention maps, we conclude that *VTTB* consistently concentrates on the area directly in front of the tool and predicts the surface of the area. We conclude from the tactile attention maps that our model can pay attention to areas where a contact deformation happens.

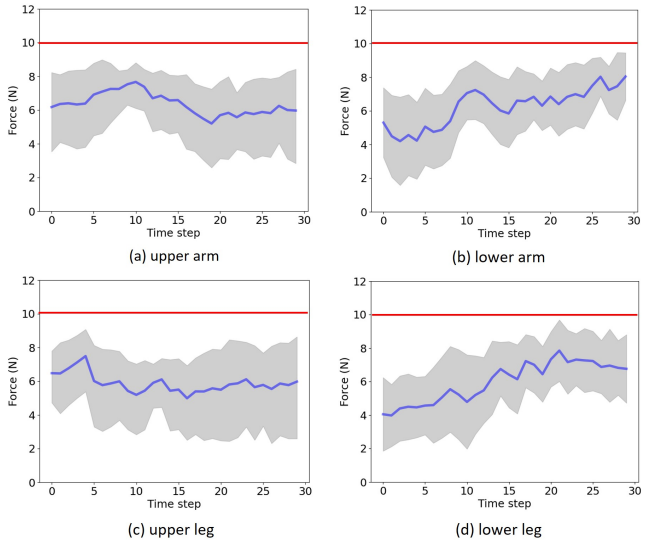

Fig. 7. The average force applied by the Baxter robot to the manikin's body parts during assistive bed bathing. The results show the average force across the test trials for four body parts, with shaded regions representing one standard deviation. We also present that the force never exceeds 10N, which is within an acceptable safe bound.

The result of the evaluation is shown in Fig. 5. We observe that the absence of depth and tactile images negatively affects task completion. One noticeable error with *No Tactile* is that the robot frequently fails to make direct contact with the manikin body. This may be caused by the reason that the robot cannot estimate whether the tool is in contact with the object surface with the lack of tactile inputs. For *No Image*, notably, the model fails to complete the task most of the time. The robot is unable to follow the contour of the body parts due to the lack of depth information.

D. Baselines Comparison

Our *VTTB* employs Transformer-based imitation learning to predict actions, as described in Section III-B. We further examine the efficacy of our model by comparing it to four imitation learning baselines:

- *BC*: the basic imitation learning baseline which clones human actions based on current observations, as in [40].
- *BC-RNN*: BC with a RNN policy which predicts the action from a temporal sequence of past observations using a Recurrent Neural Network (a 2-layer LSTM here), as in [38], [43]. *BC-RNN* is proved to be particularly effective in learning from human datasets.
- *BC-ViT*: BC with a ViT policy which learns the action using a standard ViT encoder, as in [35]. We modify the ViT encoder to enable both visual and tactile image inputs. We compare our model with it to show the efficacy of cross-attention transformer encoder.
- *BC-VTT*: BC with a VTT policy which proposes an extension to ViT encoder, as in [32]. This extension incorporates an additional contact embedding aimed at consolidating contact states, as well as an alignment embedding to align temporally latent states. We compare our model with it to show the efficacy of our design of cross-attention transformer encoder.

We train all the baselines with both visual and tactile image inputs.

Fig. 5 presents that *VTTB* outperforms all four baselines. We have found that *VTTB* generally produces the highest performance among all tested body parts. This confirms that our proposed approach with a cross-attention mechanism has strong power in the assistive bed bathing task. Qualitatively, we observe that with an attention mechanism, all of the Transformer-based policies can adapt to nonlinear surface curves while *BC* and *BC-RNN* tend to reach wrong positions where either the tool moves away from the surfaces or applies too much force to the surface. Compared to *BC-ViT* and *BC-VTT*, *VTTB* is more robust to various nonlinear curves. Even when the robot deviates from the surface due to the curves, *VTTB* can control itself back to the body part again.

We qualitatively visualize the temporal attentions of *VTTB* across both visual and tactile image inputs in Fig. 6. For each image input sequence, we show the top 4 regions with the highest attention weights at three time frames. From the attention map of the depth images, we learn that the model continues paying attention to the area directly in front of the tool, enabling the robot to predict the surface and follow the body contour. Moreover, the attention map of the tactile images shows that when the robot is experiencing a curve,

TABLE I

EVALUATION STUDY OF *VTTB* ON THREE NOVEL BODY PARTS AND TWO HUMAN SUBJECTS.

	Completed	Partially Bathed	Failed
Chest	80%	10%	10%
Waist	90%	5%	5%
Back	80%	15%	5%
Human	83.75%	12.5%	3.75%

the model is able to focus on the contact deformation area.

We also demonstrate that *VTTB* can provide bed bathing assistance while remaining an acceptable force on the body. We verify the force profile using a force/torque sensor (Robotiq FT300). Fig. 7 presents the average force the robot applied to the manikin's different body parts. We observe that the cleaning cloth is continually in contact with the body while applying less than 10N of force on average.

E. Ability of Generalization

Further, we explore whether our proposed approach has the ability to generalize to novel body parts previously unseen in the demonstration dataset. We evaluate *VTTB* on the waist, chest, and back, each with 20 repeated trials. As shown in Table I, the policy can achieve a success rate of 90% on the waist and 80% on the chest and back, which indicates that the policy generalizes well across geometric variations.

We also conduct a preliminary study with two human subjects (i.e., the authors) to explore the bathing performance of our proposed approach on real human body (see Fig. 8). We evaluate *VTTB* with the human subjects on four body parts: upper arm, lower arm, upper leg, and lower leg. For each body part, we repeat the experiment 10 times with randomized body poses and start poses. Table I shows that *VTTB* successfully completes over 80% of the experiments, indicating our method to be a promising approach for real-world bathing care. Additionally, we notice that the model demonstrates slight variations in performance between the two human subjects. This disparity may be attributed to individual differences in body surface, underscoring the need for a more extensive human study in future research.

V. DISCUSSION AND CONCLUSION

This research introduces a visuo-tactile Transformer-based imitation learning approach that provides safe in-contact bathing assistance. The fusion of visual sensing and tactile sensing explicitly models strong in-contact local relations to facilitate bathing performance. With a cross-attention mechanism, the robot learns to focus on bathing features and predicts bathing trajectories adapting to contact surfaces. Our empirical results show how our learned model outperforms imitation learning baselines and generalizes across multiple surface geometries with novel body poses and body parts and to two real human subjects.

Although our approach has demonstrated substantial success in learning robust visuo-tactile servoing control policies,

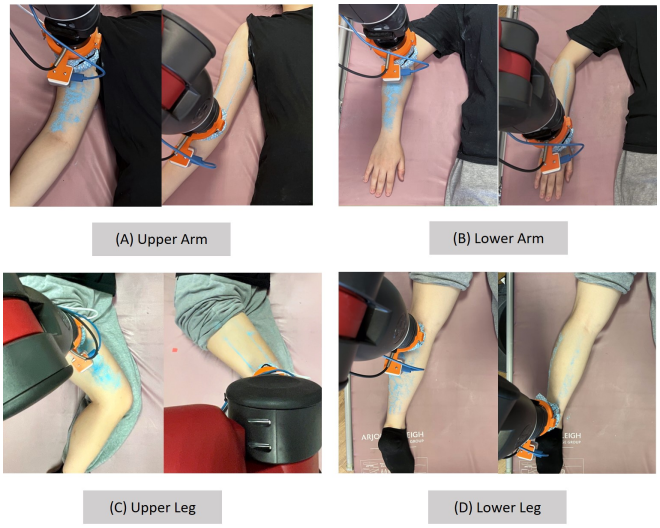

Fig. 8. Snapshots of Baxter robot providing bathing assistance to a real human for four body parts. The example trials are also provided on our project website.

several limitations guide our future work. Firstly, our current dataset is limited to a manikin, whose biomechanics differ from the biomechanics of real people [44]. Further extensions include collecting offline demonstrations directly on real humans with pre-trained policies. Furthermore, our current model primarily focuses on bathing the upper surface of body parts. By incorporating 3D coverage path planning techniques [45], we aim to further bath the body parts with comprehensive coverage. With a head-mounted camera, the robot can also learn a visual classifier to determine whether the dirt has been removed, enabling policy execution until all the dirt has been bathed. In addition, Baxter's substantial size and relative bulkiness necessitate considerable manual effort to manipulate its arms for demonstrations. In future work, we plan to transit to more maneuverable robots, such as Kinova or Franka Emika, to streamline the demonstration process.

REFERENCES

[1] T. M. Gill, Z. Guo, and H. G. Allore, "The epidemiology of bathing disability in older persons," *Journal of the American Geriatrics Society*, vol. 54, no. 10, pp. 1524–1530, 2006.

[2] J. C. Millán-Calenti, J. Tubío, S. Pita-Fernández, I. González-Abraldes, T. Lorenzo, T. Fernández-Arruty, and A. Maseda, "Prevalence of functional disability in activities of daily living (adl), instrumental activities of daily living (iadl) and associated factors, as predictors of morbidity and mortality," *Archives of gerontology and geriatrics*, vol. 50, no. 3, pp. 306–310, 2010.

[3] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8320–8329.

[4] R. Ye, W. Xu, H. Fu, R. K. Jenamani, V. Nguyen, C. Lu, K. Dimitropoulou, and T. Bhattacharjee, "Rcare world: A human-centric simulation world for caregiving robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 33–40.

[5] A. C. Dometios, Y. Zhou, X. S. Papageorgiou, C. S. Tzafestas, and T. Asfour, "Vision-based online adaptation of motion primitives to dynamic surfaces: application to an interactive robotic wiping task," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1410–1417, 2018.

[6] A. Zlatintsi, I. Rodomagoulakis, P. Koutras, A. Dometios, V. Pitsikalis, C. S. Tzafestas, and P. Maragos, "Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3171–3175.

[7] A. G. Perry, P. A. Potter, and W. Ostendorf, *Clinical nursing skills and techniques*. Elsevier Health Sciences, 2013.

[8] C.-H. King, T. L. Chen, A. Jain, and C. C. Kemp, "Towards an assistive robot that autonomously performs bed baths for patient hygiene," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 319–324.

[9] Z. Erickson, H. M. Clever, V. Gangaram, G. Turk, C. K. Liu, and C. C. Kemp, "Multidimensional capacitive sensing for robot-assisted dressing and bathing," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2019, pp. 224–231.

[10] I. Huang, D. Chow, and R. Bajcsy, "Soft tactile contour following for robot-assisted wiping and bathing," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7797–7802.

[11] R. Madan, S. Valdez, D. Kim, S. Fang, L. Zhong, D. T. Virtue, and T. Bhattacharjee, "Rabbit: A robot-assisted bed bathing system with multimodal perception and integrated compliance," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 472–481.

[12] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: learning multimodal manipulation policies with deep reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8298–8304.

[13] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[14] W. Liang, F. Fang, C. Acar, W. Q. Toh, Y. Sun, Q. Xu, and Y. Wu, "Visuo-tactile feedback-based robot manipulation for object packing," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1151–1158, 2023.

[15] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," in *Conference on Robot Learning*. PMLR, 2023, pp. 3142–3166.

[16] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, "General in-hand object rotation with vision and touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 2549–2564.

[17] R. Okumura, N. Nishio, and T. Taniguchi, "Tactile-sensitive newtonianvae for high-accuracy industrial connector insertion," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4625–4631.

[18] P. Falco, S. Lu, A. Cirillo, C. Natale, S. Pirozzi, and D. Lee, "Cross-modal visuo-tactile object recognition using robotic active exploration," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5273–5280.

[19] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2722–2727.

[20] J. Kerr, H. Huang, A. Wilcox, R. Hoque, J. Ichnowski, R. Calandra, and K. Goldberg, "Self-supervised visuo-tactile pretraining to locate and follow garment features," in *Robotics: Science and Systems*, 2023.

[21] F. Zhang and Y. Demiris, "Visual-tactile learning of garment unfolding for robot-assisted dressing," *IEEE Robotics and Automation Letters*, 2023.

[22] E. Aertbeliën and J. De Schutter, "etasl/etc: A constraint-based task specification language and robot controller using expression graphs," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1540–1546.

[23] C. A. V. Perico, J. De Schutter, and E. Aertbeliën, "Combining imitation learning with constraint-based task specification and control," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1892–1899, 2019.

[24] Z. Deng, J. Mi, Z. Chen, L. Einig, C. Zou, and J. Zhang, "Learning human compliant behavior from demonstration for force-based robot manipulation," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2016, pp. 319–324.

[25] M. Tykal, A. Montebelli, and V. Kyrki, "Incrementally assisted kinesthetic teaching for programming by demonstration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 205–212.

[26] L. Roveda, M. Magni, M. Cantoni, D. Piga, and G. Bucca, "Assembly task learning and optimization through human's demonstration and machine learning," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 1852–1859.

[27] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," in *Conference on robot learning*. PMLR, 2017, pp. 357–368.

[28] Y. Li, J. Song, and S. Ermon, "Infogail: interpretable imitation learning from visual demonstrations," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3815–3825.

[29] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[30] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Object-centric imitation learning for vision-based robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1199–1210.

[31] S. Dasari and A. Gupta, "Transformers for one-shot visual imitation," in *Conference on Robot Learning*. PMLR, 2021, pp. 2071–2084.

[32] Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 2026–2040.

[33] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 785–799.

[34] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[38] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1678–1690.

[39] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[40] M. Bain and C. Sammut, "A framework for behavioural cloning." in *Machine Intelligence 15*, 1995, pp. 103–129.

[41] P. Kingma Diederik and J. B. Adam, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015.

[42] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[43] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[44] Z. Erickson, Y. Gu, and C. C. Kemp, "Assistive vr gym: Interactions with real people to improve virtual assistive robots," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 299–306.

[45] C. S. Tan, R. Mohd-Mokhtar, and M. R. Arshad, "A comprehensive review of coverage path planning in robotics using classical and heuristic algorithms," *IEEE Access*, vol. 9, pp. 119 310–119 342, 2021.