APPENDIX

A. Minimization of KL divergence

We now show that minimizing the KL divergence (6) is equivalent to maximizing the ELBO as in (7) and (8). Recall the learning objective of CCHP:

$$\min_{\phi} D_{\mathsf{KL}}(q_\phi(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) \mid\mid p(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C)) \tag{13}$$

First, we have

$$p(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) = \frac{p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, z)p(z)}{p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C)}. \tag{14}$$

Plugging in the objective of (13) and expand, we have

$$\begin{aligned} D_{\mathsf{KL}} = \mathbb{E}_{z\sim q_\phi(z|\boldsymbol{x}_T,\boldsymbol{y}_T,\boldsymbol{x}_C,\boldsymbol{y}_C)}[&\log q_\phi(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) \\ &- \log p(z) - \log p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, z) \\ &+ \log p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C)] \geq 0 \end{aligned} \tag{15}$$

Rearranging, we have

$$\begin{aligned} &\log p(\boldsymbol{y}_T \mid \boldsymbol{x}_C, \boldsymbol{y}_C, \boldsymbol{x}_T) \\ \geq &\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{x}_T,\boldsymbol{y}_T,\boldsymbol{x}_C,\boldsymbol{y}_C)}[\log p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, z) \\ &- \log q_\phi(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) + \log p(z)] \\ = &\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{x}_T,\boldsymbol{y}_T,\boldsymbol{x}_C,\boldsymbol{y}_C)}[\log p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, z)] \\ &- D_{\mathsf{KL}}(q_\phi(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) \mid\mid p(z)) \triangleq \mathsf{ELBO} \end{aligned} \tag{16}$$

Plug (15) and (16) into (13), we have

$$\begin{aligned} &\min_{\phi} D_{\mathsf{KL}}(q_\phi(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) \mid\mid p(z \mid \boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{x}_C, \boldsymbol{y}_C)) \\ \equiv &\min_{\phi} -\mathsf{ELBO} + \log p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C) \\ \equiv &\max_{\phi} \mathsf{ELBO} \end{aligned} \tag{17}$$

since $\log p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C)$ is intractable and does not depend on ϕ. Parameterizing the likelihood $p(\boldsymbol{y}_T \mid \boldsymbol{x}_T, \boldsymbol{x}_C, \boldsymbol{y}_C, z)$ in (16), we arrive at (7) and (8). ■

B. Train Test Data Collection

We collect user demonstrations in short clips, each lasting for around 5 seconds. We collect in total 72 clips from each user. 24 of the clips contain translation or rotation only (*type 1*) with the remaining 48 clips contain both translation and rotation (*type 2*). Clips of each type above are split evenly among D_{train} and $D_{\mathsf{test,in-sample}}$ for $U_{\mathsf{in-sample}}$. Among the 24 type 2 clips assigned to $D_{\mathsf{test,in-sample}}$, half contains motions that also appear in D_{train}.

C. Model Architecture

The configurations of different MLP modules introduced in section IV-C are given in table III. Each configuration starts with input size and ending with output size. A ReLU activation is applied after input layer and every hidden layer. The hidden state size H in section IV-C1 is 128.

Module	Description	Configuration
f_{hand}	Hand feature (section IV-C1)	[320, 128, 64, 32]
f_ϕ	Latent posterior predictor in eq. (11)	[128, 128, 128, 128, 128] and two heads [128, 32]
f_y	Operation prediction in section IV-C2	[156, 128, 64, 2], one for each motion dimension

Table III: MLP configurations for CCHP neural network implementation.

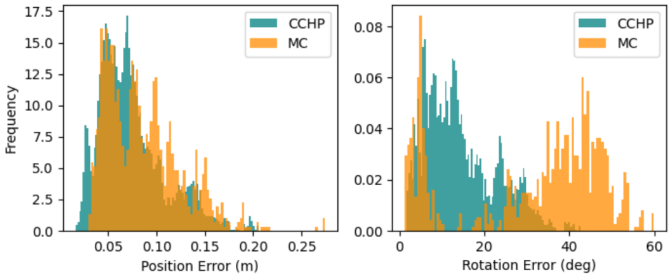

Figure 9: Distribution of top-10% cumulative errors under the "Matching" setting for CCHP and MC.

D. Model Training

We train each model using the Adam optimizer with learning rate $lr = 5e-4$. During training, command finger velocities $\boldsymbol{x}$ are perturbed with Gaussian noises $\epsilon \sim N(0, 1e-6)$. We train our models for a total of 738 epochs with batch size 32, which took around 5 hours using an Nvidia GeForce RTX 2080Ti GPU with an Intel i9-9940X CPU.

E. Cumulative Errors of CCHP and MC

To view how the improvement in velocity error (see section V-C) lead to significant difference in practice, we plot top 10% cumulative errors within 5 seconds interaction under the "Matching" setting (see table I) in fig. 9. We see that the rotation errors concentrates at 10 degrees with CCHP and at 40 degrees with motion cloning. Such gap would lead to significantly different robot behaviors for human-robot interaction.