
X2C: A Dataset Featuring Nuanced Facial Expressions for Realistic Humanoid Imitation

Peizhen Li¹, Longbing Cao¹, Xiao-Ming Wu², Runze Yang¹, Xiaohan Yu¹

¹Macquarie University ²Sun Yat-sen University

Abstract

The ability to imitate realistic facial expressions is essential for humanoid robots engaged in affective human–robot communication. However, the lack of datasets containing diverse humanoid facial expressions with proper annotations hinders progress in realistic humanoid facial expression imitation. To address these challenges, we introduce **X2C** (Anything to Control), a dataset featuring nuanced facial expressions for realistic humanoid imitation. With **X2C**, we contribute: 1) a high-quality, high-diversity, large-scale dataset comprising 100,000 (image, control value) pairs. Each image depicts a humanoid robot displaying a diverse range of facial expressions, annotated with 30 control values representing the ground-truth expression configuration; 2) **X2CNet**, a novel human-to-humanoid facial expression imitation framework that learns the correspondence between nuanced humanoid expressions and their underlying control values from **X2C**. It enables facial expression imitation in the wild for different human performers, providing a baseline for the imitation task, showcasing the potential value of our dataset; 3) real-world demonstrations on a physical humanoid robot, highlighting its capability to advance realistic humanoid facial expression imitation.

Code & Data: <https://lipzh5.github.io/X2CNet/>

1 Introduction

Rapid progress in humanoid robotics has been observed across various domains such as reception [1], education [2, 3] and healthcare [4, 5], where humanoid robots engage in affective communication with humans. These robots are increasingly deployed to interact socially, provide assistance, or support learning, making their ability to express emotions a key factor in fostering trust, empathy, and engagement [6, 7, 8, 9, 10]. Effectively delivering affective information is crucial for enhancing user experiences and ensuring meaningful interactions between humans and robots. Consequently, there has been growing emphasis on enabling humanoid robots to imitate realistic and authentic facial expressions, as facial expressions play a central role in conveying emotional cues [11, 12].

Despite recent advances in humanoid facial expression imitation [13, 14], the fidelity of humanoid facial expressions—especially the fine-grained, nuanced emotional cues—remains difficult to guarantee due to the scarcity of data required to learn emotional subtleties and guide informed on-robot execution. Existing datasets of humanoid facial expressions for the imitation task (Smile [13], Co-expression [14]) are typically small in size, lack sufficient data diversity (e.g., they do not include asymmetric facial expressions), and the emotional nuances that can be learned are limited by low annotation dimensionality (see Table 1). Their annotation accuracy is not guaranteed as the dataset collection relies on facial landmark predictions [15], which introduce prediction errors.

To bridge the gap, we introduce **X2C**, a new resource for realistic humanoid facial expression imitation. It consists of 100,000 (image, control value) pairs, with each image depicting a humanoid robot displaying a diverse range of nuanced facial expressions (Figure 1). Each image is annotated with 30 numerical control values representing the ground-truth expression configuration. These

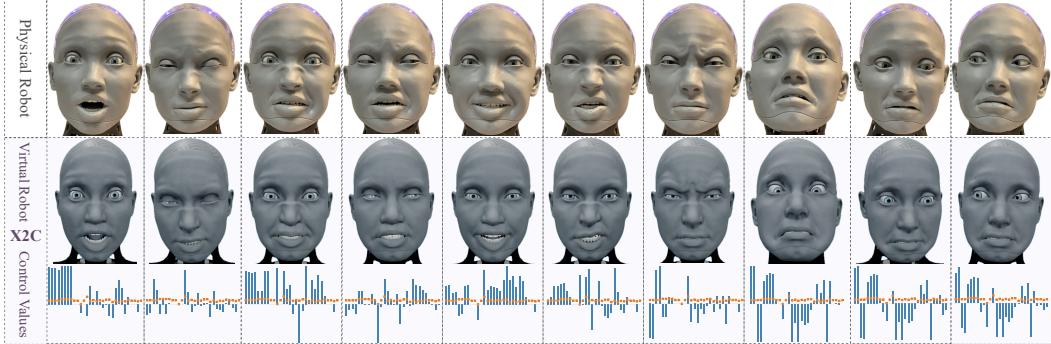


Figure 1: **Demonstration of X2C dataset examples.** Each example in the X2C dataset consists of: (1) an image depicting the virtual robot, shown in the middle; and (2) the corresponding control values, visualized at the bottom. In these visualizations, the height of each **blue bar** represents the magnitude of the corresponding value, while the **orange dots** indicate the values in the neutral state.

Table 1: Summary statistics of existing datasets for realistic humanoid facial expression imitation.

Dataset	Dataset Characteristics				Dataset Quality		
	Size	Asymmetric Expressions	Input Dimensionality	Annotation Dimensionality	Annotation Accuracy	Data Alignment	Data Diversity
X2C	100,000	✓	512 × 512 × 3	30	★★★★★	✓	★★★★★
Smile [13]	15,000	✗	480 × 320 × 3	10	★★★★	✓	★★★
Coexpression [14]	1000	✗	113 × 2	11	★★★★	✓	★★★

control values can be used to drive the physical humanoid robot to reproduce the expression shown in the image. **Asymmetric facial expressions** [16] are also included in our dataset (e.g., the 2nd column in Figure 1) to simulate the human-like behavior and encourage diversity. To facilitate understanding of the relationship between the physical and virtual robots, images of the physical robot are also included at the top of Figure 1. The physical robot and its virtual counterpart share the same set of controls, ensuring consistent facial expressions across both platforms. In principle, facial expressions should **be independent** of the skin type (or identity) of the humanoid robot [17, 18]. Therefore, issues such as sim-to-real gap [19] do not raise, since control values encode emotional nuances and serve as a bridge between the virtual and physical robots.

Images in the dataset are extracted from videos of the humanoid robot performing facial expression animations. These animations are manually curated by volunteers from different birth countries and of different genders (including females and males), to help eliminate potential biases related to cultural background and gender. The dataset includes basic facial expressions (e.g., surprise, joy, and sadness [20]) at varying intensities (e.g., the last two in Figure 1 show fear with subtle differences in gaze and head pose), as well as complex expressions that may not fit neatly into basic emotion categories (e.g., the 4th column in Figure 1). This is done purposefully to ensure broad **expression coverage**. To ensure **consistency**, the upper-body pose of the robot remains fixed during video recording, and all images are resized to a uniform resolution. To guarantee the **uniqueness** of the dataset, we apply structural similarity checks to the captured images to identify and remove near-duplicate frames, resulting in 100,000 retained images with diverse humanoid facial expressions for annotation. To ensure annotation **accuracy**, we formulate interpolation equations based on parameters specified in the animation files. For control value sampling, i.e., given a timestamp s , precise continuous control values will be provided by the equations. The timestep for sampling both control values and images is set to 0.05 seconds, enabling perfect alignment between images and their corresponding control annotations [21, 22, 23], thereby ensuring the dataset quality. Details of the control values sampling and annotation process are provided in Section 2.3. Dataset characteristics such as size (number of samples), input dimensionality, and annotation dimensionality, as well as dataset quality metrics such as annotation accuracy, data alignment and data diversity [24] are summarized in Table 1. To our knowledge, X2C is the first high-quality, high-diversity, large-scale dataset featuring nuanced humanoid facial expressions specifically designed for realistic humanoid imitation. Equipped with X2C, we introduce X2CNet, a novel framework for realistic humanoid

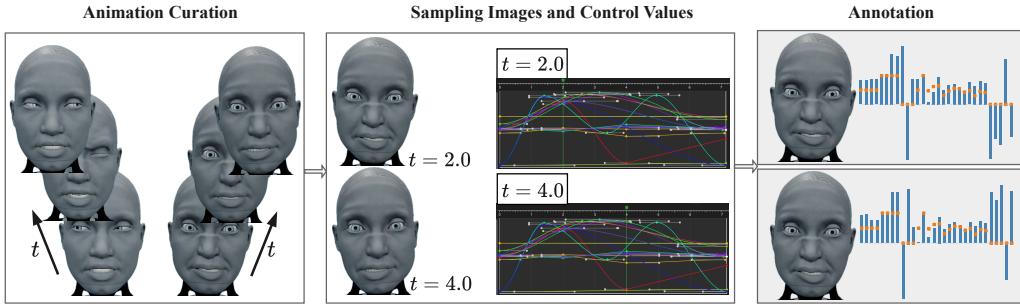


Figure 2: The pipeline for dataset collection. We first curate humanoid facial expression animations covering all basic emotions and beyond. Images and their corresponding control values are then sampled at the same timestamps (e.g., if an image is sampled at $t = 2.0$, its control value annotation is also sampled at $t = 2.0$) to obtain the temporally aligned pairs. See Section 2.3 for more details.

facial expression imitation. It decomposes the humanoid learning process into two stages: in the first stage, the expression dynamics is captured through a motion transfer module; in the second stage, the correspondence between nuanced humanoid facial expression and their underlying control values is learned via large-scale training using **X2C**. This framework enables facial expression imitation in the wild for different human performers. An overview of **X2CNet** is provided in Figure 5.

After introducing **X2C** (Section 2), we demonstrate its value in enabling human-to-humanoid learning by presenting an imitation framework, **X2CNet** (Section 3), and further validate its potential through real-world experiments on physical humanoid robots performing nuanced facial expression imitation tasks. Beyond imitation, **X2C** opens up new research avenues, including human-like motion generation, facial expression evaluation for robots [25, 26] and the development of expressive humanoid robots for affective human–robot interaction [27].

2 The X2C Dataset

X2C has been made publicly available and provides images of nuanced humanoid facial expressions along with ground-truth control value annotations. A comparison of the data characteristics and quality of **X2C** against existing datasets for realistic humanoid facial expression imitation is presented in Table 1.

In the following sections, we first introduce the humanoid robot and the control value preliminaries (Section 2.1). Before providing an overview of the dataset (Section 2.4), we detail the dataset collection process: volunteers were invited to curate facial expression animations for the humanoid robot and record videos (Section 2.2); we then sampled images from videos and calculated control values using mathematical equations (Section 2.3) to construct the (image, control value) pairs. The pipeline of dataset collection is provided in Figure 2. 10 volunteers were recruited from the student population at Macquarie University, all of whom were over 18 years old and provided informed consent. To minimize possible biases related to different background, we purposefully selected volunteers with different birth countries, including undergraduate and PhD students of different genders. Dataset collection ran from 15nd November 2024 to 15nd January 2025¹

2.1 The Humanoid Robot and Control Values

The humanoid robot employed for dataset collection is called Ameca (Figure A2), which features 32 Degrees of Freedom (DoFs) including facial actuators and head/neck movements (Figure A3). The higher number of facial DoFs—compared with most existing humanoid robots [28, 29, 30, 31)—allows for finer-grained and more nuanced facial expressions. There are 30 control values associated with these DoFs, which are responsible for driving actuators located at different expression-relevant control units, including the brows, lids, gaze, nose, mouth, head and neck.

¹Ethics approval, data collection, and analysis was led by researchers from Macquarie University.

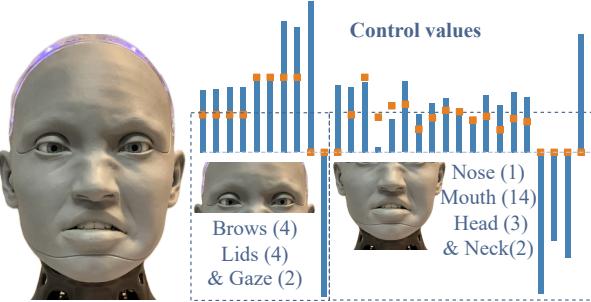


Figure 3: **An illustration of the correspondence between control values and control units.** In the control value visualization, the first 4 values control the brow movements, the next 4 control eyelid motions, and so on for the other units.

Table 2: Names of the controls and the corresponding ranges of their values.

Jaw Pitch [0, 1]	Jaw Yaw [0, 1]	Lip Bottom Curl [0, 1]	Lip Bottom Depress Left [0, 1]	Lip Bottom Depress Middle [0, 1]	Lip Bottom Depress Right [0, 1]
Lip Corner Raise Left [0, 1]	Lip Corner Raise Right [0, 1]	Lip Corner Stretch Left [0, 1]	Lip Corner Stretch Right [0, 1]	Lip Top Curl [0, 1]	Lip Top Raise Left [0, 1]
Lip Top Raise Middle [0, 1]	Lip Top Raise Right [0, 1]	Nose Wrinkle [0, 1]	Brow Inner Left [0, 1]	Brow Inner Right [0, 1]	Brow Outer Left [0, 1]
Brow Outer Right [0, 1]	Eyelid Lower Left [-1, 2]	Eyelid Lower Right [-1, 2]	Eyelid Upper Left [-1, 2]	Eyelid Upper Right [-1, 2]	Gaze Target Phi [-2.3, 2.3]
Gaze Target Theta [-1.1, 1.1]	Head Pitch [-0.5, 0.3]	Head Roll [-0.3, 0.3]	HeadYaw [-0.5, 0.5]	Neck Pitch [-0.3, 0.5]	Neck Roll [-0.3, 0.3]

The distribution of control values across these units is visualized in Figure 3, and the value ranges for each control are provided in Table 2. Note that the number of control values and DoFs differ. This is because the gaze controls (Gaze Target Phi and Gaze Target Theta) drive the left and right eyes symmetrically, which reflects a human-like design, as humans exhibit conjugate gaze [32].

2.2 Humanoid Expression Animations

Environment Preparation Exploration of the robot’s expression space requires arbitrary combinations of different control values (each sampled within its legal range), which is necessary to curate a high-diversity dataset. Some combinations may lead to rare facial expressions that most humans cannot display (Figure A1), and frequently driving the robot with such control values can cause irreversible damage to the robot’s silicon skin, leading to expensive repairs. For safety and to minimize mechanical wear on the physical robot, we chose to conduct dataset collection in a simulation environment where a virtual counterpart of the physical robot is available. Given the same control values, the virtual robot displays the same facial expressions as the physical one. There will be no issues such as the sim-to-real gap [19] since facial expressions should be disentangled from the robot’s skin (or identity) [17, 18]. Images from the physical robot and its virtual counterpart are equivalent in the sense that they have the expression embedding in robot’s action space, represented by control values. Details of the data collection environment are provided in the Supplemental Material. This environment can be accessed simultaneously by multiple certified accounts, which accelerates the data collection process.

Expression Animation Curation After 30 hours of training, volunteers became familiar with rigging the robot in the aforementioned environment and acquired the necessary prerequisites for animators. They were then tasked with creating key-framed animations [33, 34], each defined by a sequence of critical frames that capture the most significant humanoid expressions at key moments in time, along with corresponding interpolation methods [35]. Intermediate frames (in-betweens) are then interpolated to produce smooth expression animations. The currently available interpolation methods include **Cubic Bézier**, **Linear** and **Step** interpolations. **Cubic Bézier** interpolation provides a smooth transition by blending four control points P_0, P_1, P_2, P_3 , where P_0 and P_3 are the start and end points, and P_1, P_2 are control points. The interpolation is defined over the normalized parameter

$u \in [0, 1]$ [36]:

$$I(u) = (1 - u)^3 P_0 + 3(1 - u)^2 u P_1 + 3(1 - u)u^2 P_2 + u^3 P_3, \quad u \in [0, 1]. \quad (1)$$

Linear interpolation creates a straight-line transition between two keyframe values P_0 and P_1 defined at times t_0 and t_1 , respectively. The interpolation function is given by [37]:

$$I(t) = P_0 + (P_1 - P_0) \cdot \frac{t - t_0}{t_1 - t_0}, \quad t \in [t_0, t_1]. \quad (2)$$

Step interpolation holds a constant value until the next keyframe. For two keyframes at times t_0 and t_1 , with values P_0 and P_1 , the step interpolation is defined as [38]:

$$I(t) = \begin{cases} P_0, & \text{if } t \in [t_0, t_1) \\ P_1, & \text{if } t = t_1 \end{cases} \quad (3)$$

Volunteers could specify the interpolation method and value for each control at a critical moment to form keyframes. These values were purposefully selected to sweep the full range of each control as much as possible, ensuring broad coverage of the expression space and promoting diversity in the dataset. The resulting 560 animations (with durations ranging from 1 to 15 seconds) then underwent a subsequent processing stage.

2.3 Sampling and Annotation

We filmed videos of all curated animations using the same device (MacBook Air, 2020) and under consistent configurations (i.e., frame rate, and upper-body pose of the robot) to ensure data consistency. Images were then sampled at a constant timestep of $s = 0.05$ seconds and resized to a uniform resolution of 512×512 pixels. To ensure data uniqueness, the resulting images were processed with a structural similarity check to identify and remove near-duplicate frames if their similarity exceeded a threshold ($\theta = 0.99$). Given two image patches x and y , the Structural Similarity Index (SSIM) is computed as follows [39]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (4)$$

where C_1 and C_2 are stability constants, defined as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, with $L = 255$, $K_1 = 0.01$ and $K_2 = 0.03$. μ_x , μ_y denote the mean of x and y respectively; σ_x^2 and σ_y^2 are their variances; and σ_{xy} is the covariance between x and y . To obtain the control value annotations, we retrieve the keyframes and interpolation parameters from the animation metadata, formulate the corresponding interpolation equations (as shown in equations (1) to (3)) and sample precise control values at the same timestamps used to extract images from the animations. This ensures annotation accuracy and temporal alignment between the images and their annotations.

The overall dataset collection pipeline is summarized in Figure 2, where we provide a visualization for control curve sampling at two timestamps (i.e., $t = 2.0$ and $t = 4.0$).

2.4 Dataset Overview

Table 3: Summary statistics of the 30 control values. The names of the controls are abbreviated using the first capital letter of each word. Full names are provided in Table 2.

	JP	JY	LBC	LBDL	LBDM	LBDR	LCRL	LCRR	LCSL	LCSR	LTC	LTRL	LTRM	LTRR	NW
μ	0.635	0.519	0.544	0.565	0.465	0.560	0.355	0.542	0.646	0.403	0.543	0.649	0.540	0.620	0.191
σ	0.366	0.190	0.123	0.085	0.087	0.064	0.129	0.087	0.119	0.095	0.153	0.134	0.182	0.140	0.287
V_{\max}	1.000	1.000	0.991	0.750	0.759	0.822	0.802	0.976	1.000	0.991	1.000	1.000	1.000	1.000	1.000
V_{\min}	0.000	0.000	0.000	0.148	0.000	0.185	0.000	0.250	0.295	0.300	0.000	0.475	0.300	0.409	0.000
V_{neu}	1.000	0.500	0.460	0.560	0.430	0.540	0.470	0.620	0.640	0.310	0.410	0.480	0.300	0.450	0.000
	BIL	BIR	BOL	BOR	ELL	ELR	EUL	EUR	GTP	GTT	HP	HR	HY	NP	NR
μ	0.605	0.655	0.613	0.598	1.110	1.060	0.989	0.978	0.074	0.045	0.004	0.005	-0.002	0.008	0.002
σ	0.236	0.233	0.211	0.216	0.521	0.518	0.311	0.331	0.311	0.113	0.036	0.021	0.048	0.024	0.009
V_{\max}	1.000	1.000	1.000	1.000	2.000	2.000	2.000	2.269	1.082	0.328	0.300	0.355	0.204	0.084	
V_{\min}	0.000	0.000	0.000	0.000	-1.000	-1.000	-1.000	-2.269	-0.826	-0.371	-0.173	-0.413	-0.104	-0.158	
V_{neu}	0.500	0.500	0.500	0.500	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

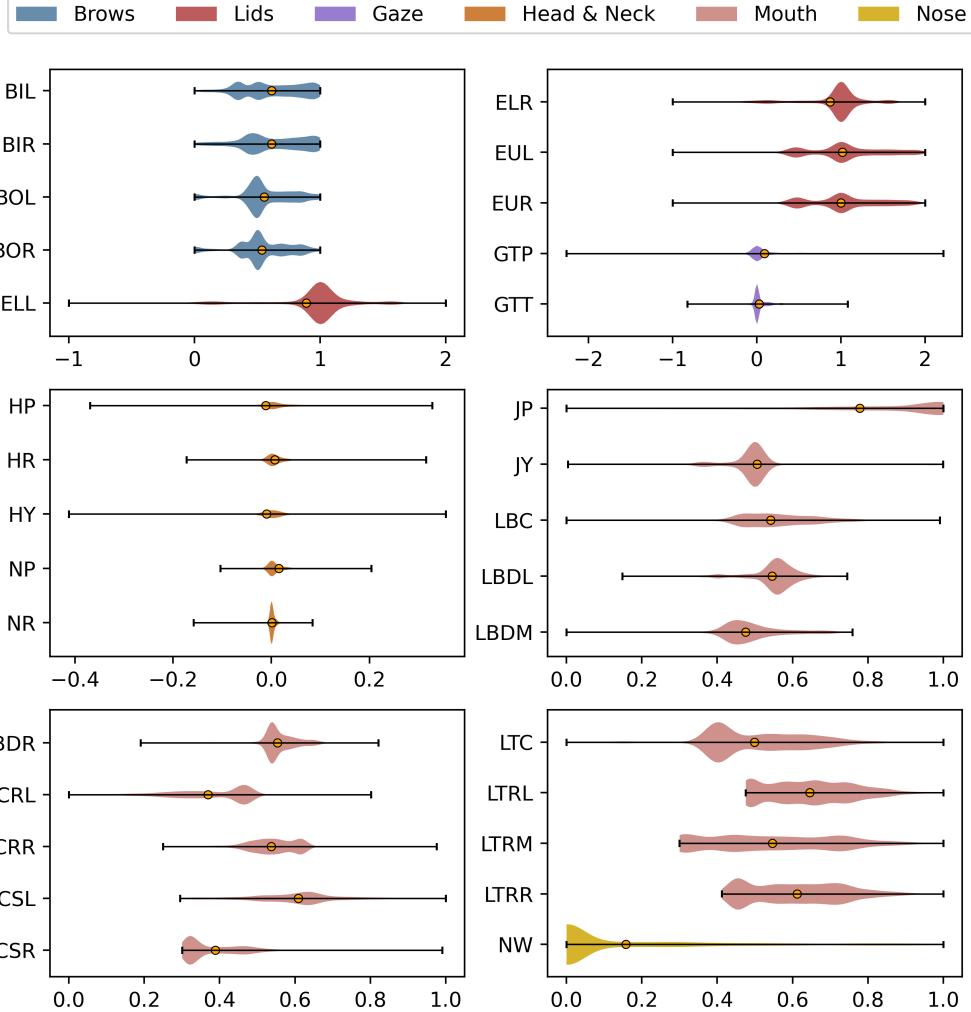


Figure 4: **Value distributions of 30 controls.** Controls for different expression-relevant units are indicated by different colors. The naming convention for controls is the same as in Table 2.

Examples from **X2C** are provided in Figure 1, where the humanoid robot displays a wide range of facial expressions. Some expressions may not fit neatly into basic emotion categories (e.g., the 4th), some of them can be categorized into the same emotion class (fear) but with different intensities (e.g., the last two) and some of them contain asymmetric facial units (e.g., eyelids of the 2nd). We name the dataset **X2C** (Anything to Control) because it consists of pairs of emotion-aware representations (humanoid facial expressions, in this case) and their corresponding control values (used for guiding on-robot execution). We aim to expand this dataset by incorporating fine-grained emotion labels in our future work.

Characteristics and Quality Comparing to existing datasets [13, 14] (Table 1) for humanoid facial expression imitation, **X2C** offers a larger scale and higher-dimensional annotations, enabling finer-grained robot control. Unlike prior works that rely on tools like MediaPipe [40] to estimate facial landmarks—introducing potential errors—our control values are analytically calculated via interpolation functions, ensuring precise annotations that are perfectly aligned with images at each timestamp. Distinctly, our dataset includes asymmetric facial expressions [41], which are closer to natural human behavior. This not only enhances diversity but also improves the expressive capacity of the robot. We report summary statistics for each of the 30 control values in Table 3, including the mean (μ), standard deviations (σ), minimum (V_{\min}) and maximum (V_{\max}). Values in the neutral state (V_{neu}) are also provided for reference. As shown, there are noticeable deviations between μ and V_{neu} .

across most controls, and for many controls (e.g., JP, JY, LBC), the dataset samples span nearly the full achievable range as indicated in Table 2, suggesting high-diversity. Note that we intentionally avoid sampling extreme values for certain controls because: 1) They could cause irreversible damage to the robot (e.g., excessive head or neck movement such as HP, HR, HY, NP, NR may lead to mechanical wear), and 2) Such expressions are physiologically implausible for humans—for instance, humans typically cannot fully hide their irises (GTP, GTT) or shape their lips into extreme forms like a ‘W’ or ‘V’ (see Figure A1). For clarity, a visualization of the value distributions for all 30 controls is provided in Figure 4.

3 The X2CNet Framework

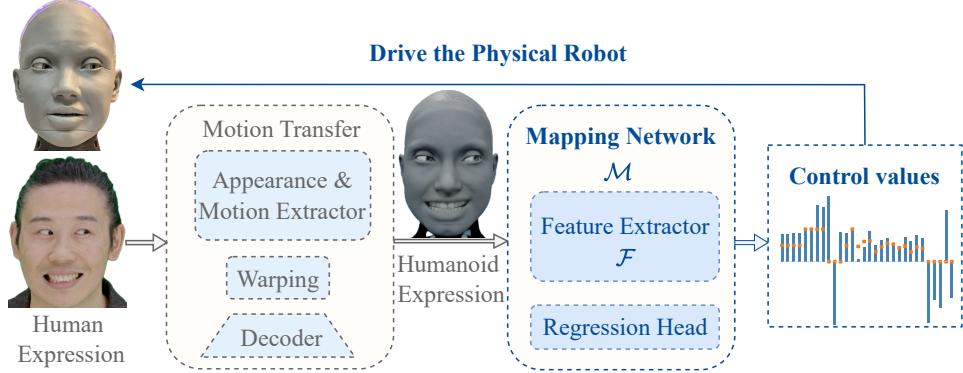


Figure 5: An overview of X2CNet, the proposed imitation framework. The first module captures facial expression subtleties from humans, while the mapping network learns the correspondence between various humanoid expressions and their underlying control values using the **X2C** dataset.

3.1 Motivation and Design

The objective of this framework is to demonstrate the value of our dataset in advancing nuanced humanoid facial expression imitation. It should be able to learn from **X2C** and enable the humanoid robot to realistically mimic human facial expressions. There are two key requirements for such a framework: 1) It must be capable of capturing subtle expression dynamics from humans. 2) It must output low-level action commands that interface with the robot’s control system [42, 43, 44]. To meet the first requirement, we employ a motion transfer technique [45, 46, 47, 48, 49] to warp the humanoid face according to human expressions with emotional nuances in image space. To satisfy the second requirement, we learn the correspondence between nuanced humanoid facial expressions and the underlying control values using a mapping network trained on the **X2C** dataset.

X2CNet is thus composed of two modules (Figure 5) and outputs 30 continuous control values that encode subtle movements of expression-relevant control units. Fine-grained control, together with a delicately designed humanoid face featuring 32 DoFs, makes realistic facial expression imitation possible. From an implementation perspective, we adopt LivePortrait [48] as the motion transfer module, pretrained on a large corpus of high quality portrait data [50, 51, 52, 53]. As shown in Figure 5, it consists of an appearance extractor, a motion extractor, a warping module, and a generator. The generated humanoid face—now expressing the human emotion—is then fed into a mapping network (denoted by \mathcal{M}), which consists of a feature extractor (denoted by \mathcal{F}) and a regression head. \mathcal{F} is implemented using a ResNet18 backbone [54] while the regression head is a multilayer perceptron with two hidden layers and ReLU activations [55].

3.2 Experiments on X2C

We split **X2C** into training and test sets, using 80% of them for training. We use AdamW as the optimizer with a weight decay of 0.05 and apply a *cosine schedule with warmup* as the learning rate scheduler, with an initial learning rate of 1e-3. The model is trained using the Huber loss with

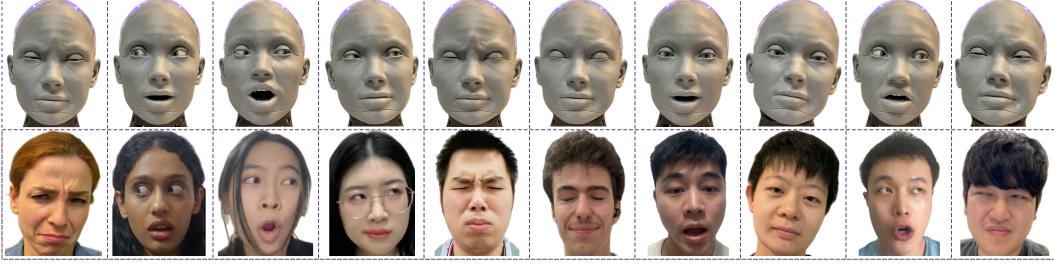


Figure 6: **Examples of realistic humanoid imitation.** Different individuals express a wide range of facial expressions, with nuances reflected in features such as frown, gaze direction, eye openness, nose wrinkles, mouth openness, and so on. These nuanced human facial expressions extend beyond canonical emotions and can be regarded as either blends of different canonical emotions or as a single emotion with varying intensities. The humanoid robot mimics every detail, resulting in a realistic imitation.

a threshold value of $\delta = 0.01$. Throughout training, the batch size is set to 128, and the model is trained for 100 epochs. All experiments are conducted on a single RTX 4090 GPU.

Control value prediction errors are evaluated on the test set using mean absolute error (MAE) as the performance measure. To assess the effectiveness of the mapping network, we compare our method against three baselines. The first baseline randomly samples each control value independently from a uniform distribution (RC). The second baseline randomly selects samples directly from the training set (RT) [56, 57]. While both involve random selection, they follow different strategies. The third baseline adopts the model architecture from [14], predicting control values based on facial landmarks (LMKC). Quantitative results are summarized in Table 4. As shown, our method outperforms all three baselines on the test set consisting of 20,000 samples, achieving lower mean errors and smaller standard deviations.

Ablation studies are conducted on the feature extractor \mathcal{F} within the mapping network. The MAEs for control value prediction, along with corresponding statistical analyses are reported in Table 5. Among the three CNN-based feature extractors—EfficientNet-B0, ResNet18, and VGG16—VGG16 achieves the best performance, followed by the transformer-based ViT-B/16. While ResNet18 performs slightly worse than both, it is significantly more lightweight and computationally efficient.

We compute the the mean absolute error (MAE) between the predicted and ground-truth control values, and conduct a detailed statistical analysis, including the calculation of the standard deviation (SD), standard error of the mean (SEM), and 95% Confidence Interval (CI).

We study alternative structures of the feature extractor \mathcal{F} within the mapping network by replacing it with the backbone of EfficientNet-B0 (EN-B0) [58], VGG16 [59] and Vision Transformer (ViT-B/16) [60]. All of them are adapted from official implementations and initialized with pretrained weights.

4 Real-World Demonstrations

To meet the requirements of real-world demonstrations, the robot must be capable of imitating a wide range of nuanced facial expressions from diverse human performers in the wild. To this end, we particularly recruited 20 human performers from 5 different countries, including both males and females. As shown in Figure 6, They exhibit a variety of facial contours, skin tones, and hairstyles. Their facial expressions were captured under different lighting conditions, and some performers appear with accessories such as glasses (4th) and earphones (6th). Performers were

Table 4: Comparison results based on MAE and corresponding statistical analysis.

Method	MAE ↓	SD ↓	SEM ↓	95% CI
RC	0.8951	0.7217	0.0051	[0.8851, 0.9051]
RT	1.0629	0.8904	0.0063	[1.0505, 1.0752]
LMKC	0.1602	0.3402	0.0024	[0.1555, 0.1629]
OURS	0.0114	0.0650	0.0005	[0.0105, 0.0123]

Table 5: An ablation study on the feature extractor.

\mathcal{F}	MAE ↓	SD ↓	SEM ↓	95%CI
EN-B0	0.0151	0.0636	0.0004	[0.0142, 0.0159]
VGG16	0.0107	0.0642	0.0005	[0.0098, 0.0116]
ViT-B/16	0.0111	0.0641	0.0005	[0.0103, 0.0120]
ResNet18	0.0114	0.0650	0.0005	[0.0105, 0.0123]

instructed to go beyond canonical expressions by incorporating various subtleties such as frowning (1st), gaze direction (2nd), and neck movement (8th). As illustrated in Figure 6, the humanoid robot successfully mimics most of these nuanced expressions despite its hardware constraints. These results validate the effectiveness of the imitation framework and, more importantly, highlight the value of **X2C**—our high-diversity, high-quality, large-scale dataset for humanoid learning. More real-world demonstrations can be found on our project website: <https://lipzh5.github.io/X2CNet/>.

5 Related Work

Facial Expressions for Affective Human-Robot Interaction Facial expressions has been proved an indispensable mode of affective communication [11, 6] and numerous studies have examined the important role of facial expressions in affective human-robot interaction (HRI) [61, 62, 63, 12, 64]. However, some of them focus on human facial expression analysis, neglecting the emotionally intelligent behavior on robot’s face, where the robots may only display limited categories of emotional signals (such as LED indicators) on their faces [65, 66, 67, 68]. The robots often fail to convey the nuances of emotions, leading to reduced user engagement and trust in HRI. By providing a large-scale, annotated humanoid facial expression dataset, **X2C** pave the path for researchers aiming at improving the robots’ facial expressiveness for HRI.

Humanoid Facial Expressions Imitation Although recent studies have paid more attention to the expression display on robots’ faces [69, 70, 71, 72], only a limited set of facial expressions are covered, which limits the expressiveness of the robot. Although efforts have been paid on nuanced facial expression imitation recently [73, 14, 13, 74, 29], there is a lack of public available resources for accessing advanced, delicate humanoid face, benchmarking different models on this task. To bridge the gap, we introduce **X2C**, the first high-quality, high-diversity, large-scale dataset featuring nuanced humanoid facial expressions with precise control value annotations for realistic humanoid imitation.

6 Limitations, Discussions and Conclusions

Limitations and Future Work While our dataset collection includes volunteers from multiple countries and genders, cultural biases may still be present, potentially influencing the interpretation or design of facial expressions. In the future, we plan to expand the sample population for recruiting animation creators and to further diversify the dataset. Although our current dataset includes facial expressions from only a single humanoid robot, the data collection pipeline and the proposed imitation framework are designed to generalize to other humanoid platforms with different degrees of freedom (DoFs). Future work will also focus on extending **X2C** with fine-grained emotion labels to enable more precise supervision for the imitation task.

Ethical Considerations and Societal Impacts All human participants involved in the real-world experiments provided informed consent, with consent forms included in the Supplemental Material. No harmful information about participants is released, and all data use follows ethical research guidelines. The dataset could be misused for deceptive, manipulative, or surveillance-related purposes, such as impersonation or unauthorized identity mimicry. We strongly discourage such applications and advocate for responsible, ethical AI use. Positively, the dataset has the potential to empower emotionally intelligent robots for socially beneficial applications, including elderly care, autism therapy, and education, by enabling more expressive and relatable interactions. However, overly human-like robots may cause users—especially vulnerable individuals—to form emotional attachments or unrealistic expectations, possibly leading to confusion or psychological discomfort. These risks must be carefully managed through transparent system design and user education.

Conclusions To equip the humanoid robot with the ability to realistically imitate nuanced human facial expressions, we make the following contributions: 1) we introduce the **X2C** (Anything to Control)—a high-quality, high-diversity, large-scale datasets featuring nuanced humanoid facial expressions with precise control value annotations; 2) we propose **X2CNet**, a novel framework for human-to-humanoid expression imitation; 3) we provide real-world demonstrations on the physical robot to validate the effectiveness of our method, and the potential of our dataset in advancing realistic humanoid facial expression imitation.

References

- [1] Ruth Maria Stock. Emotion transfer from frontline social robots to human customers during service encounters: Testing an artificial emotional contagion model. 2016.
- [2] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, 2004.
- [3] Fumihide Tanaka, Aaron Cicourel, and Javier R Movellan. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, 2007.
- [4] Hideki Kozima, Cocoro Nakagawa, and Yuriko Yasuda. Interactive robots for communication-care: A case-study in autism therapy. In *ROMAN*, 2005.
- [5] T Esubalew, Uttama Lahiri, Amy R Swanson, Julie A Crittendon, Zachary E Warren, Nilanjan Sarkar, et al. A step towards developing adaptive robot-mediated intervention architecture (aria) for children with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2012.
- [6] Niyati Rawal and Ruth Maria Stock-Homburg. Facial emotion expressions in human–robot interaction: A survey. *International Journal of Social Robotics*, 2022.
- [7] Lara Toledo Cordeiro Ottoni and Jés de Jesus Fiais Cerqueira. A systematic review of human–robot interaction: the use of emotions and the evaluation of their performance. *International Journal of Social Robotics*, 2024.
- [8] Yoonhyuk Jung, Eunae Cho, and Seongcheol Kim. Users’ affective and cognitive responses to humanoid robots in different expertise service contexts. *Cyberpsychology, Behavior, and Social Networking*, 2021.
- [9] Faruk Seyitoğlu and Stanislav Ivanov. Robots and emotional intelligence: A thematic analysis. *Technology in Society*, 2024.
- [10] Shane Saunderson and Goldie Nejat. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 2019.
- [11] Albert Mehrabian. Communication without words. In *Communication theory*. 2017.
- [12] Peizhen Li, Longbing Cao, Xiao-Ming Wu, Xiaohan Yu, and Runze Yang. Ugotme: An embodied system for affective human-robot interaction. *arXiv preprint arXiv:2410.18373*, 2024.
- [13] Boyuan Chen, Yuhang Hu, Lianfeng Li, Sara Cummings, and Hod Lipson. Smile like you mean it: Driving animatronic robotic face with learned models. In *ICRA*, 2021.
- [14] Yuhang Hu, Boyuan Chen, Jiong Lin, Yunzhe Wang, Yingke Wang, Cameron Mehlman, and Hod Lipson. Human-robot facial coexpression. *Science Robotics*, 2024.
- [15] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016.
- [16] William E Rinn. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 1984.
- [17] Hanwei Liu, Rudong An, Zhimeng Zhang, Bowen Ma, Wei Zhang, Yan Song, Yujing Hu, Wei Chen, and Yu Ding. Norface: Improving facial expression analysis by identity normalization. In *ECCV*, 2024.
- [18] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *CVPR*, 2021.
- [19] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *ICRA*, 2018.

- [20] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 1992.
- [21] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.
- [22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.
- [23] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.
- [24] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2021.
- [25] Michael J Gielniak, C Karen Liu, and Andrea L Thomaz. Generating human-like motion for robots. *The International journal of robotics research*, 2013.
- [26] Christian Becker-Asano and Hiroshi Ishiguro. Evaluating facial displays of emotion for the android robot geminoid f. In *WACI*. IEEE, 2011.
- [27] Cynthia Breazeal. Emotion and sociable humanoid robots. *International journal of human-computer studies*, 2003.
- [28] Yanghai Zhang, Changyi Liu, Keting Fu, Wenbin Zhou, Qingdu Li, and Jianwei Zhang. Fabg: End-to-end imitation learning for embodied affective human-robot interaction. *arXiv preprint arXiv:2503.01363*, 2025.
- [29] Jiayan Li, Honghao Lyu, Nan Zhang, Haiteng Wu, and Geng Yang. Design and realization of a multi-dof robotic head for affective humanoid facial expression imitation. In *ICIRA*, 2023.
- [30] Zanwar Faraj, Mert Selamet, Carlos Morales, Patricio Torres, Maimuna Hossain, Boyuan Chen, and Hod Lipson. Facially expressive humanoid robotic face. *HardwareX*, 2021.
- [31] Matthias Kerzel, Erik Strahl, Sven Magg, Nicolás Navarro-Guerrero, Stefan Heinrich, and Stefan Wermter. Nico—neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction. In *RO-MAN*, 2017.
- [32] Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. The role of fixational eye movements in visual perception. *Nature reviews neuroscience*, 2004.
- [33] Alla Safanova, Jessica K Hodgins, and Nancy S Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics*, 2004.
- [34] C Karen Liu and Zoran Popović. Synthesis of complex dynamic character motion from simple animations. *ACM Transactions on Graphics*, 2002.
- [35] Rick Parent. *Computer animation: algorithms and techniques*. Newnes, 2012.
- [36] Pierre Bézier. Numerical control-mathematics and applications. *Translated by AR Forrest*, 1972.
- [37] James D Foley. *Computer graphics: principles and practice*. Addison-Wesley Professional, 1996.
- [38] Watt Alan and Watt Mark. Advanced animation and rendering techniques. *Theory and Practice Wokingham*, 1992.
- [39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004.
- [40] Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

- [41] Rotem Kowner. Laterality in facial expressions and its effect on attributions of emotion and personality: a reconsideration. *Neuropsychologia*, 1995.
- [42] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.
- [43] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *HRI*, 2016.
- [44] Mark W Spong, Seth Hutchinson, and M Vidyasagar. Robot modeling and control. *John Wiley &*, 2020.
- [45] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019.
- [46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019.
- [47] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *NeurIPS*, 2022.
- [48] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.
- [49] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021.
- [50] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
- [51] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.
- [52] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 2018.
- [53] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *NeurIPS*, 2021.
- [54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.
- [55] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *CoRL*, 2022.
- [56] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *NeurIPS*, 2016.
- [57] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.
- [58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.
- [59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [60] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

- [61] Maartje MA De Graaf, Somaya Ben Allouch, and Jan van Dijk. Long-term acceptance of social robots in domestic environments: Insights from a user’s perspective. In *AAAI spring symposia*, 2016.
- [62] Monica N Nicolescu and Maja J Mataric. Learning and interacting in human-robot domains. *IEEE Transactions on Systems, man, and Cybernetics-part A: Systems and Humans*, 2001.
- [63] Céline Ray, Francesco Mondada, and Roland Siegwart. What do people expect from robots? In *IROS*, 2008.
- [64] Longbing Cao. Ai robots and humanoid ai: Review, perspectives and directions. *arXiv preprint arXiv:2405.15775*, 2024.
- [65] Pablo Barros, Cornelius Weber, and Stefan Wermter. Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. In *Humanoids*, 2015.
- [66] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE CAA J. Autom. Sinica*, 2017.
- [67] David O Johnson, Raymond H Cuijpers, and David van der Pol. Imitating human emotions with artificial facial expressions. *International Journal of Social Robotics*, 2013.
- [68] Diego R Faria, Mario Vieira, Fernanda CC Faria, and Cristiano Premebida. Affective facial expressions recognition for human-robot interaction. In *ROMAN*, 2017.
- [69] Ali Meghdari, Saeed Bagheri Shouraki, Alireza Siamy, and Azadeh Shariati. The real-time facial imitation by a social humanoid robot. In *ICROM*, 2016.
- [70] F Cid, José Augusto Prado, Pablo Manzano, Pablo Bustos, and Pedro Núñez. Imitation system for humanoid robotics heads. *J. Phys. Agents*, 2013.
- [71] Shuzhi Sam Ge, Chen Wang, and Chang Chieh Hang. Facial expression imitation in human robot interaction. In *ROMAN*, 2008.
- [72] Niyati Rawal, Dorothea Koert, Cigdem Turan, Kristian Kersting, Jan Peters, and Ruth Stock-Homburg. Exgennet: Learning to generate robotic facial expression using facial expression recognition. *Frontiers in Robotics and AI*, 2022.
- [73] Victor Nikhil Antony, Maia Stiber, and Chien-Ming Huang. Xpress: A system for dynamic, context-aware robot facial expressions using language models. *arXiv preprint arXiv:2503.00283*, 2025.
- [74] Xiaofeng Liu, Rongrong Ni, Biao Yang, Siyang Song, and Angelo Cangelosi. Unlocking human-like facial expressions in humanoid robots: A novel approach for action unit driven facial expression disentangled synthesis. *IEEE Transactions on Robotics*, 2024.

Appendix

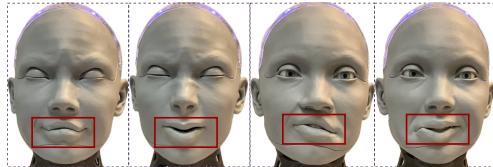


Figure A1: Examples of facial expressions that are physiologically implausible for humans ('W' shape, 'V' shape, and two asymmetric cases).

Some combinations of control values may lead to facial expressions that are physically implausible for humans. We present several examples in Figure A1 with a focus on the mouth. Both symmetric (left two) and asymmetric (right two) cases are provided.



Figure A2: An image of the humanoid robot used for dataset collection and experiments. (Image source: <https://engineeredarts.com/robot/ameca/>).

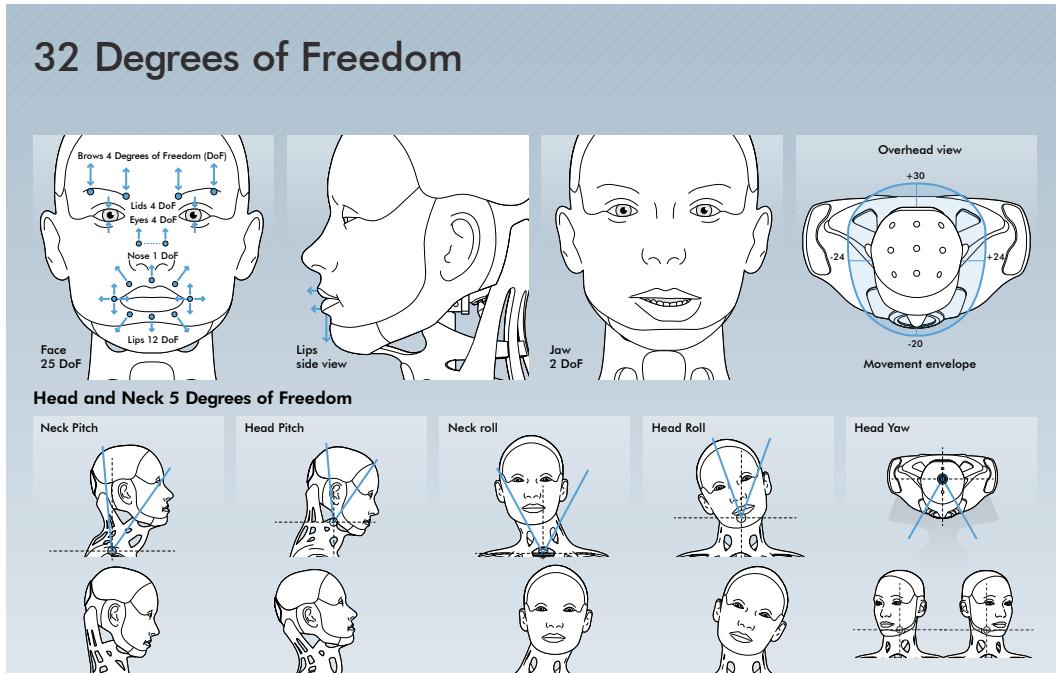


Figure A3: Demonstration of the 32 Degrees of Freedom (DoFs) on the robot's face (Figure source: <https://engineeredarts.com/robot/ameca/>).

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The paper has 3 core sections. We describe the contributions of the dataset in Section 2. Then we present the proposed framework and real-world demonstrations in Section 3 and Section 4 respectively.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We summarize limitations in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide details for the framework design and real-world demonstration settings in Section 3 and Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release all code needed to reproduce results on our Github:<https://github.com/lipzh5/X2CNet>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We describe the experimental setting including data splits and hyperparameters in detail in Section 3.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We provide statistical analysis on the prediction errors by calculating standard deviations, standard error of the mean, 95% confidence interval. The results are described in Section 3.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We use a single RTX 4090 GPU for model training and inference as described in Section 3.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in this paper fully conforms to the NeurIPS Code of Ethics. All data were collected with appropriate consent and no harmful information about the human participants is released.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have provided a discussion of the societal impacts of our work in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly cited the original creators of all existing assets used in this work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduce a new dataset. The dataset and its dataset card are publicly available at: <https://huggingface.co/datasets/Peizhen/X2C>.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[Yes\]](#)

Justification: We have included the consent form provided to the human volunteers involved in the real-world experiments in the Supplemental Material.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[Yes\]](#)

Justification: We have received ethics approval from Macquarie University.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorosity, or originality of the research, declaration is not required.

Answer: [\[NA\]](#)

Justification: The core methods for dataset collection and framework design do no involve LLMs as any important, original or non-standard component.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.