2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Improving Robotic Imitation Learning with Predicted Facial Motion Using Transformers

Yitong Li and Fumio Kanehiro, *Member, IEEE*

***Abstract*— This study proposes a Transformer-based approach with cross-attention for predicting human facial movements in face-related robotic control tasks and integrating these predictions into an imitation learning framework. A dataset of human facial videos was constructed, and landmarks were extracted using the MediaPipe framework. Three prediction methods were compared, and the cross-attention model achieved the best performance in both landmark localization accuracy and image quality. In imitation learning experiments, facial motion trajectories sampled from real human data trajectories were used, and the success rate increased from 42% to 60% and ultimately to 74% when predicted landmarks were incorporated. Additionally, varying the prediction horizon affected task completion time, with the 2-frame horizon achieving the fastest completion. These results demonstrate that incorporating predicted facial motion can significantly enhance robotic control performance in dynamic human-robot interaction scenarios.**

Keywords— Transformer, Cross-attention, Facial Motion Prediction, Imitation Learning, Robotic Feeding

I. Introduction

Recent advancements in Human-Robot Interaction (HRI) have enabled robotic arms to be increasingly deployed in healthcare and assistive applications for the elderly and disabled [1]. Among these applications, face-related tasks—including feeding, providing water, and face wiping—are particularly critical because they require precise operations in close proximity to the human face [2]. Such interactions demand both high task accuracy and strict safety guarantees [3].

However, most existing robotic arm control methods rely primarily on static facial landmark information for motion planning. This reliance presents significant limitations in dynamic scenarios, as human faces may exhibit sudden motions or posture changes during interaction [4]. Failing to perceive or predict these movements in time can lead to operational errors or even safety risks [5].

To address these challenges, we propose predicting future facial movements to allow robotic arms to proactively plan their actions in face-related tasks, thereby improving task success rates. Specifically, a cross-attention Transformer is employed to predict the temporal evolution of facial landmarks across multiple future frames, and these predictions are integrated into the Action Chunking with Transformers (ACT) imitation learning framework [6], enabling the robot to generate motion sequences that adapt to facial dynamics.

The main contributions of this study are summarized as follows:

1. We propose a cross-attention-based Transformer model for predicting future human facial movements, which significantly improves both landmark localization accuracy and the quality of generated images.
2. We validate the proposed approach in a robotic feeding task with motion trajectories sampled from real human data, achieving a success rate improvement from 42% to 74%.

II. Related Work

Imitation learning has been extensively studied in robotic control, demonstrating high efficiency and strong generalization capabilities, particularly in high-dimensional, continuous action spaces [7]. Common methods include Behavior Cloning [8] and Inverse Reinforcement Learning [9], which have been widely applied to robotic trajectory generation, object manipulation, and human-robot collaboration. However, these approaches typically rely on static input states, which limits their performance in dynamic human-robot interaction scenarios.

The Transformer architecture, originally developed for natural language processing, has shown remarkable success in modeling long-range dependencies through its self-attention mechanism. Owing to its sequential modeling capability, the Transformer has been increasingly applied to motion prediction [10], trajectory generation [11], and human-robot interaction [12]. Studies in human motion prediction, gesture recognition, and robotic path planning have demonstrated that Transformers can effectively capture temporal dependencies and reduce prediction errors [13][14]. In robotic path planning, their ability to model long-term temporal dependencies has been shown to improve decision-making performance [15]. Nonetheless, the application of Transformers to facial dynamics prediction, and their integration into robotic control frameworks, remains underexplored.

Facial landmark detection and tracking are fundamental for face-related robotic tasks such as feeding and face wiping. Traditional approaches often rely on Convolutional Neural Networks (CNNs) to detect facial landmarks from single images but lack the ability to model temporal continuity [16]. More recent work has explored using RNNs and Transformers to incorporate temporal information for landmark trajectory

*Research supported by JST Spring, Grant Number JPMJSP2124.

Yitong Li (corresponding author) and Fumio Kanehiro are with the Doctoral Program in Intelligent and Mechanical Interaction Systems, University of Tsukuba, and the CNRS-AIST JRL (Joint Robotics Laboratory), IRL, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8560, Japan (e-mail: li.cornel@aist.go.jp and f-kanehiro@aist.go.jp).

prediction, improving detection stability in dynamic scenarios [17][18]. However, these methods still struggle to maintain robust predictions when landmarks are temporarily occluded due to head movement.

The Action Chunking with Transformers (ACT) framework was recently proposed to learn temporal patterns of robotic action sequences using Transformers [19]. ACT has demonstrated strong performance in imitation learning, efficiently generating long-horizon action chunks [20]. Yet, research combining future facial motion prediction with the ACT framework—particularly for dynamic adaptation in face-related tasks—remains limited.

In summary, while prior studies have advanced imitation learning, Transformer-based sequence modeling, and facial landmark prediction, there is still a gap in integrating Transformer-predicted facial motion into imitation learning frameworks for adaptive control in face-related robotic tasks. This study aims to address this gap by integrating Transformer-predicted facial motion into an imitation learning framework.

III. Approach

The overall pipeline of the proposed approach is illustrated in Fig. 1. It consists of two main modules, a Transformer-based prediction model and an imitation-learning-based control model. The ACT control policy receives RGB images from three cameras along with the robot's joint states, while the Transformer module uses only the front-facing close-up camera to predict facial motion. The Transformer model outputs facial landmark coordinates, or more generally, facial motion information for several future frames. These predicted features are then fed into the control model as part of the observation input. The control model, built upon the Action Chunking with Transformers (ACT) framework, combines the predicted features with the current state of the robotic arm to generate a sequence of future actions. The following subsections describe the architecture of each module in detail.

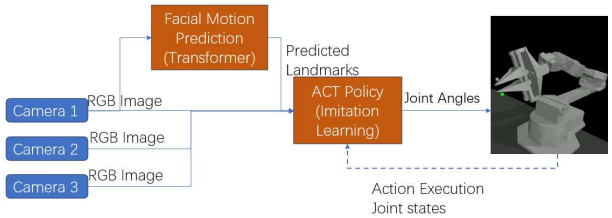

Figure 1. Overall architecture of the proposed system. Facial motion prediction provides future facial landmarks to the ACT policy, which outputs the robot trajectory.

A. Facial Motion Prediction Using Transformer

Transformers, originally proposed for natural language processing tasks, are now widely applied in computer vision and robotics due to their strong ability to model long-range dependencies through self-attention mechanisms. In this study, we leverage the sequential modeling capability of Transformers to predict the temporal evolution of facial landmarks over several future frames, making them particularly suitable for capturing dynamic variations in facial landmark trajectories.

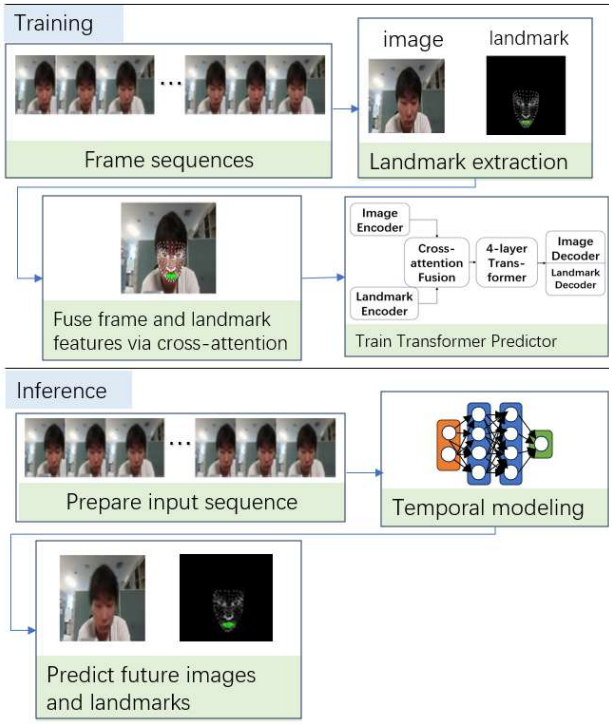

Figure 2. Training and inference pipeline of the facial motion prediction module.

To achieve this, the input to the Transformer is a video sequence consisting of multiple consecutive frames. Each frame contains an RGB image and the normalized 2D coordinates of all facial landmarks. The mouth region (e.g., landmarks around the lips) is particularly important for the control task and is highlighted in green in visualizations. In this work, we predict normalized 2D landmark coordinates (U, V), where u and v are image-space positions used for normalization.

In addition to predicting mouth landmarks, the model also predicts the corresponding RGB frames as an auxiliary supervision signal. This encourages the Transformer to learn more robust temporal features, even though only landmark predictions are ultimately used for robotic control. Moreover, predicting the full set of landmarks, rather than just the mouth center, provides richer geometric information and results in more stable motion features. The mouth center used for control can then be derived from these predicted landmarks.

A key innovation of this work is the introduction of a cross-attention mechanism to effectively integrate information from two different modalities: RGB images and facial landmarks. Instead of simply concatenating features or processing them independently, the cross-attention operation enables the model to learn the relationship between the visual content of images and the spatial geometry described by the landmarks. The overall training and inference pipeline of the facial motion prediction module is illustrated in Fig. 2.

Mathematically, let the feature matrix of the image sequence be denoted as $I \in R^{T \times D_i}$, where T is the number of frames and D_i is the feature dimension of the image embeddings. Similarly, let the landmark feature matrix be $\boldsymbol{L} \in R^{T \times D_l}$, where D_l is the feature dimension of the landmark embeddings. The cross-attention mechanism computes the attention weights between the landmarks and the image features as follows:

$$A = softmax\left(\frac{L \cdot I^T}{\sqrt{d}}\right) \quad (1)$$

where d is the dimension of the projected feature vectors used in the attention computation. These attention weights are then used to compute a combined context representation:

$$C = A \cdot I \quad (2)$$

Specifically, for each time step i, the fused feature vector c_i is given by:

$$c_i = \sum_{j=1}^{T} softmax\left(\frac{L_i I_j^T}{\sqrt{d}}\right) \cdot I_j \quad (3)$$

where L_i and I_j are the landmark and image feature vectors at time steps i and j, respectively. Note that L and I refer to the complete sequences.

The fused features $c_{i_{i=1}}^{T}$ are then passed into the Transformer, which models temporal dependencies and directly predicts the coordinates of facial landmarks for the upcoming frames.

The model is trained using the mean squared error (MSE) loss between the predicted and ground truth landmark coordinates, defined as:

$$\mathcal{L}_{MSE} = \frac{1}{T}\sum_{t=1}^{T}\|\hat{y}_t - y_t\|^2 \quad (4)$$

where $\hat{y}_t$ and y_t denote the predicted and ground truth landmark coordinates at time step t, respectively.

B. Imitation Learning Framework

At the control level, this study is based on the Action Chunking with Transformers (ACT) framework, which uses Transformer models to map system states to sequences of robotic actions from demonstration data, predicting short action chunks over several future frames.

In face-related robotic tasks, the ultimate goal is to safely and precisely interact with real human faces. However, conducting experiments directly on humans poses significant risks. Robotic operations near the human face carry the possibility of physical harm, and accidental incidents may result in legal or ethical concerns. Moreover, collecting real-world demonstration data is time-consuming, particularly when considering diverse facial dynamics, pose variations, and task scenarios.

To address these challenges, we conduct experiments in a simulated environment. Simulation ensures safety by preventing physical harm, allows rapid experimental iterations, and eliminates hardware wear, thus improving research efficiency. This approach is especially useful for evaluating new methods in the early stages of development; only after sufficient validation in simulation will the approach be considered for real-world transfer.

We further design a method to integrate Transformer-based facial motion predictions into simulation-based imitation learning. The Transformer model is trained on real-world facial video data to predict the future dynamics of facial movements. Fig. 3 illustrates the ACT-based imitation learning framework, including observation construction, policy inference, and trajectory generation.

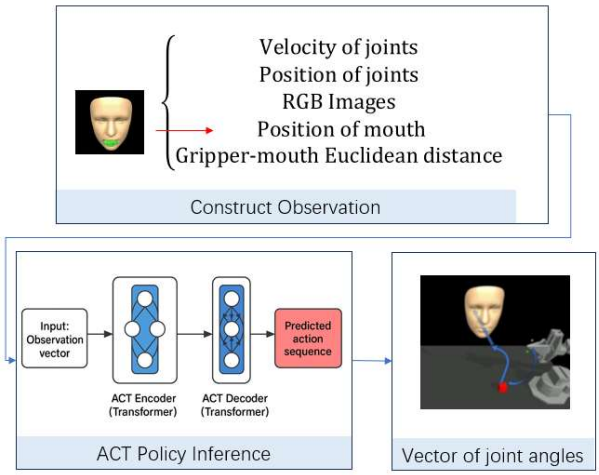

Figure 3. ACT-based imitation learning workflow from demonstration recording to trajectory output.

Although the trained Transformer model is not directly used in simulation—because the simulated face lacks realistic motion—its prediction error statistics are incorporated to improve realism. Specifically, since the simulated environment provides ground-truth mouth positions, the observations fed to the imitation learning model are perturbed using average prediction errors from real-world data. For example, if the simulated mouth position is (X, Y, Z) the perturbed observation becomes ($X + \delta x$, $Y + \delta y$, Z), where δx and δy are sampled according to the average absolute prediction errors.

By incorporating prediction uncertainties into the imitation learning process, this design allows the robot to better adapt to deviations in observed facial motion.

IV. Experiments and Results

This section presents the experiments and designs performed in this study, covering Transformer-based prediction and the ACT imitation learning experiments.

A. Facial Motion Prediction Using Transformer

The purpose of this experiment is to evaluate whether a Transformer model can effectively predict future facial movements, to assess its feasibility for subsequent robotic control tasks.

Videos of human facial movements were recorded at 30 FPS, producing approximately one hour of data covering both deliberate and natural facial behaviors. The recorded motions include head movements such as nodding, turning, and

slight positional shifts or tilts, as well as mouth dynamics such as opening and closing, subtle lip movements, and natural activities like eating, drinking, reading, and using a phone. In total, 1,000 sequences were collected, each consisting of 75 frames. Among them, 900 sequences were used for training and 100 for validation, with no overlap. The sequence length of 75 frames was chosen to provide approximately 2 s of contextual input (70 frames) and a short prediction window (5 frames), offering a balance between temporal representation and computational efficiency. Facial landmarks were extracted using the MediaPipe framework [21], generating the data for both training and evaluation.

A Transformer-based network was implemented using PyTorch to jointly predict facial images and landmarks. RGB images were resized to 256×256 px, and 468 facial landmarks were extracted per frame via MediaPipe.

The architecture consists of a convolutional image encoder, a linear landmark encoder, and a cross-attention module for fusing visual and landmark features. Fused features are then processed by a 4-layer Transformer encoder to capture temporal dependencies. Two decoders output the next five frames of images and landmarks, respectively. The network was trained using the Adam optimizer with an initial learning rate of 1e-4. The total training loss combines image reconstruction and landmark regression, both computed using mean squared error (MSE):

$$Loss_{total} = Loss_{img} + 0.5 \times Loss_{landmark} \quad (5)$$

We evaluated three approaches for Transformer-based facial motion prediction:

1. Image-only prediction: Predicting future images directly and then extracting facial landmarks from the predicted images.
2. Simple feature fusion: The second approach employed a simple fusion strategy, where both image data and landmark data were fed into a single Transformer model to simultaneously predict future images and future landmarks.
3. Cross-attention fusion (proposed): Introducing a cross-attention mechanism into the model to integrate image features and landmark features.

B. Transformer Prediction Results

Figures 4, 5, and 6 share the same layout, each displaying the prediction result of the final frame. Top-left is ground truth image; Top-right is ground truth image overlaid with ground truth landmarks; Bottom-left is image predicted by the Transformer model; Bottom-right is predicted image overlaid with predicted landmarks. Green dots represent mouth landmarks and white dots represent other facial landmarks. This visualization enables a clear comparison of image quality and landmark alignment across different approaches.

Figure 4 presents the results of Approach 1. It can be observed that the predicted image is extremely blurry, making it nearly impossible to recognize any facial details. Moreover, due to the excessive blurriness, no meaningful facial landmarks could be extracted, and thus Approach 1 failed to produce usable landmark predictions.

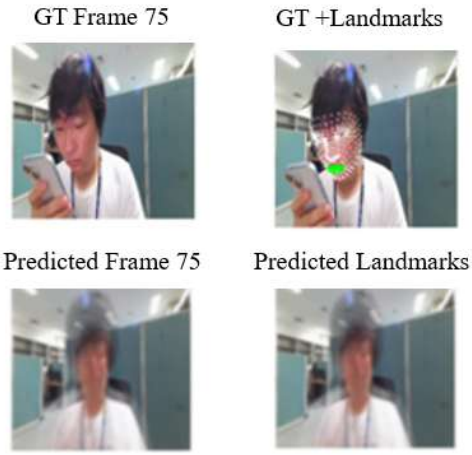

Figure 4. Visualization of prediction results for Approach 1, which directly predicts future images and then extracts facial landmarks from the predicted images.

Figure 5 shows the results of Approach 2. Although this method was able to output predicted landmarks, the predicted image remains highly blurry, and the landmarks are scattered. Notably, the green dots, which should indicate the shape of the mouth, are positioned in a disorganized manner, indicating significant errors in landmark localization.

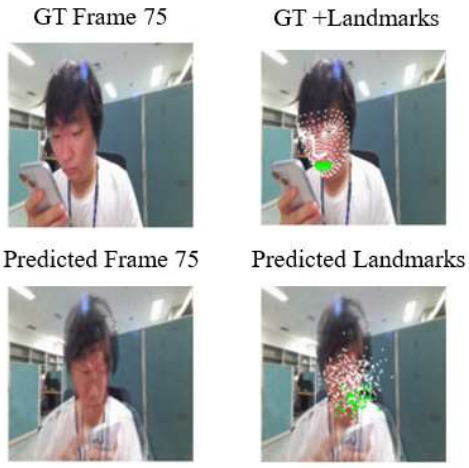

Figure 5. Visualization of prediction results for Approach 2, which employs a simple fusion strategy to simultaneously predict future images and facial landmarks from combined image and landmark features.

Figure 6 illustrates the results of Approach 3. Compared to the previous methods, the predicted image demonstrates a substantial improvement in clarity, and the predicted landmarks are more accurately concentrated and closely match the shape and position of the ground truth landmarks. This indicates that the cross-attention approach effectively enhances both image generation quality and landmark localization accuracy.

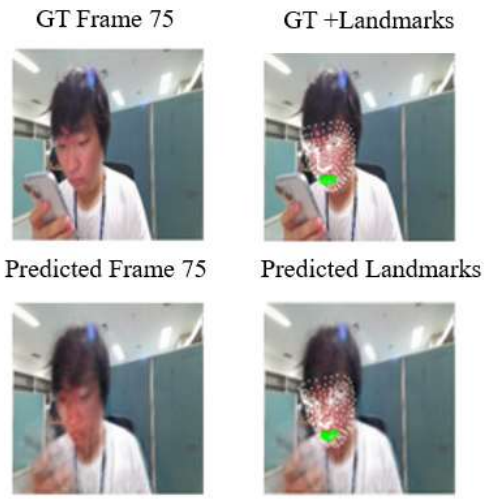

Figure 6. Visualization of prediction results for Approach 3, which introduces a cross-attention mechanism to integrate image features and landmark features for joint prediction.

TABLE I. COMPARISON OF PREDICTION PERFORMANCE

Metric	Approach		
	Approach 1	*Approach 2*	*Approach 3*
Mean ΔU	——	0.032	**0.015**
Var ΔU	——	0.061	**0.013**
Mean ΔV	——	0.036	**0.017**
Var ΔV	——	0.130	**0.025**
Average SSIM	0.662	0.818	**0.928**

TABLE I summarizes the quantitative results, comparing the prediction errors of the mouth center position and the SSIM metric. The mouth center is defined as the average of the U and V coordinates of all mouth-related landmarks where the raw pixel coordinates (u, v) are first normalized to (U, V) in [0, 1] after resizing the image to 256×256. The reported errors ΔU and ΔV represent the mean absolute differences between predicted and ground truth mouth centers, normalized by the image size (256×256). The mean values reported in TABLE I are the averages of the absolute differences over the validation set. The evaluation was conducted on a validation set comprising 100 samples , with 5 predicted frames per sequence, resulting in a total of 500 frames used for quantitative analysis. It can be observed from both the figures and the TABLE I that Approach 3 consistently outperformed Approach 2 and 1 in all evaluated aspects.

TABLE II shows the per-frame prediction performance of Approach 3. The mean error for Frame 71 is notably lower than that of the subsequent frames, while the remaining frames exhibit no clear trend in error variation. SSIM values remain consistently high across all frames.

TABLE II. PER-FRAME PREDICTION PERFORMANCE OF APPROACH 3

Metric	Predicted Frame				
	Frame 71	*Frame 72*	*Frame 73*	*Frame 74*	*Frame 75*
Mean ΔU	**0.010**	0.019	0.013	0.020	0.014
Var ΔU	**0.010**	0.014	0.014	0.013	0.014
Mean ΔV	**0.013**	0.021	0.015	0.017	0.017
Var ΔV	**0.021**	0.027	0.026	0.025	0.027
Average SSIM	0.926	0.927	0.928	0.928	**0.929**

C. *ACT Imitation Learning under Varying Observation Strategies*

To evaluate how different types of observation inputs affect imitation learning performance, we designed and compared multiple observation strategies for the feeding task, including baseline (which only includes the robot's joint states, gripper pose, and 3 RGB camera inputs), current mouth position information, future prediction, and varying the prediction horizon.

All imitation learning experiments were conducted in the RoboManipBaselines simulation environment [22]. To implement the proposed framework, we integrate the predicted facial motion into the imitation learning setup. The robot receives visual and landmark observations and attempts to move a cube toward the predicted mouth region.

In the experimental environment, two VX300s robotic arms are placed; however, only the right arm is used to execute the task. The scene also includes a red cube placed on the table and a human face model located at a random position. Figure 7 shows the setup used.

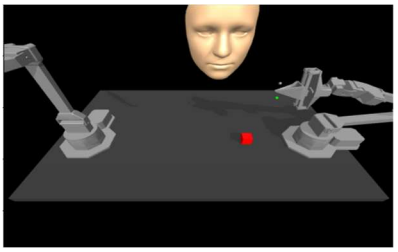

Figure 7. Visualization of the simulated imitation learning environment in RoboManipBaselines.

In this study, a scripted policy was used to generate demonstration trajectories for the feeding task. The face trajectory is directly obtained from recorded demonstration videos: a random starting time is selected, and the following 12s of 3D mouth-center positions are used as the trajectory without any modification, so that the simulated facial motion faithfully follows real human motion. For each observation configuration, 50 trajectories were collected independently to ensure reproducibility, under identical conditions with a fixed random seed, ensuring consistent facial motions and cube positions across configurations. Algorithm 1 summarizes the step-by-step procedure of this scripted policy.

Algorithm 1 SCRIPTED POLICY FOR FEEDING TASK

// Initialization
Initialize robot_arm at initial pose
Initialize time_step as dt = 0.02s
Initialize face_model at random initial position (base position + random offset within ±0.05 m along X, Y, and Z)
// Define facial motion trajectory
Randomly select a start time t0 within the pre-recorded trajectory and extract a 6-second segment
target_position = initial_face_position + (trajectory(t0+6) – trajectory(t0))
// Phase 1: Approach and grasp cube
Move robot_arm toward cube position
Open gripper
Close gripper to grasp the cube
// Phase 2: Move toward mouth (face moves along face_velocity)
t ← 0
while (|gripper.x - mouth_position.x| >= 0.03 m OR
|gripper.y - mouth_position.y| >= 0.01 m OR
|gripper.z - mouth_position.z| >= 0.01 m AND
t <= 6s) do:
move_velocity = (current_mouth_position - current_gripper_position) / (6 - t)
Move face_model according to pre-sampled trajectory
Move gripper by move_velocity * dt
t ← t + dt
end while
// Phase 3: Insert cube into mouth and release
Open gripper to release the cube
Retract robot_arm from mouth to initial position (face fixed)

Each configuration was evaluated over 50 independent trials, and the success rate was calculated as the proportion of trials in which the robot successfully passed the cube through the mouth region, as defined by the spatial condition in Algorithm 1.

The ACT framework is used as the imitation learning policy. TABLE III summarizes the hyperparameters used for training. To ensure fair comparisons, all hyperparameters remain consistent across different observation settings.

TABLE III. HYPERPARAMETERS FOR TRAINING THE ACT POLICY

Hyperparameter	Value
KL divergence weight	10
Chunk size (action sequence)	100
Transformer hidden dimension	512
Feedforward network dimension	3200
Batch size	8
Learning rate	1E-5
Training epochs	2000

To systematically compare the impact of different observation configurations, we defined the following experimental groups: (1) Baseline: Without any landmark or prediction information. (2) Current Only: Including only the current mouth position and gripper–mouth distance. (3) Current + Prediction: Adding future predicted mouth positions with varying horizons (1 to 5 frames).

Error Injection: For (3), two variants were tested—one with accurate data and one with random perturbations based on the average prediction errors of the Transformer model. The normalized coordinate prediction errors ΔU, ΔV can be converted into spatial errors δx, δy as follows:

$$max_{\delta x} = 2Ztan(FOV_x/2)\Delta U \quad (6)$$

$$max_{\delta y} = 2Ztan(FOV_y/2)\Delta V \quad (7)$$

where Z is the current depth (distance to the camera. An Acer FHD Camera (built-in laptop webcam) was used as the visual input, with a 75° horizontal FOV and 47° vertical FOV. In our recorded videos, the camera – mouth distance was approximately Z≈0.5m. Under this setting, the corresponding spatial deviations are approximately $|\delta x| \lesssim 0.012$m and $|\delta y| \lesssim 0.007$m.

In the baseline setting, the observation space includes: the positions and velocities of the robot's six joints; the gripper's position and velocity; and three RGB camera inputs (top, angled, and front-close views), each with a resolution of 640×480. Figure 8 shows example images captured from the three RGB cameras (top, angled, and front-close views) used in the baseline observation setting.

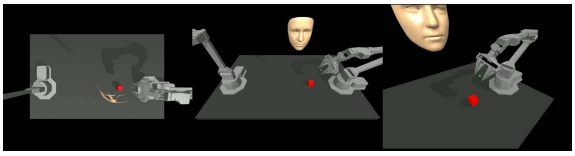

Figure 8. Example views from the three RGB cameras (top, angled, and front-close).

In the experimental setting with only current mouth position information, compared to the baseline, the observation additionally includes the current coordinates of the mouth region, as well as the Euclidean distance between the gripper and this point.

In the experimental setting with future prediction, compared to the baseline, the observation additionally includes a time-ordered sequence of predicted mouth positions in the format of $(t_0, x_0, y_0, z_0, t_1, x_1, y_1, z_1, \dots)$, where (x_k, y_k, z_k) represents the predicted mouth positions at time k, and t_k is the corresponding time encoding. The observation also includes the Euclidean distances between the gripper and each predicted point, which are directly derived from the predicted landmark sequence.

The goal of this experiment is to evaluate whether incorporating prediction-based features can improve the success rate of imitation learning in a facial targeting task.

D. ACT Imitation Learning Results

The ACT imitation learning experiments were conducted using the observation strategies defined in the previous section. Each configuration was evaluated on the robotic feeding task, and the results are summarized in TABLE IV.

TABLE IV. SUCCESS RATES UNDER VARYING OBSERVATION STRATEGIES AND PREDICTION HORIZONS

Observation Type	Horizon (frames)	Success Rate	
Baseline	None	42%	
Current Only	0	60%	
Current + prediction		With Error	Without Error
	1	58%	66%
	2	72%	**74%**
	3	68%	72%
	4	62%	68%
	5	54%	60%

In this study, the baseline success rate was 42%, and providing only the current mouth position increased it to 60%. The highest success rate was obtained with a 2-frame or 3-frame prediction horizon without error, reaching 74%. When prediction errors are introduced (with error), the trend of success rate variation with respect to the frame horizon remains consistent with the without error case. However, the success rates are consistently lower than those of the corresponding without error experiments.

To further investigate the effect of different prediction horizons, TABLE V reports mean task completion times under the Current + Prediction configuration. Task completion time is defined as the duration until the gripper reaches the predicted mouth position. The results indicate that completion time vary with the prediction horizon. Notably, the 2-frame horizon achieved the shortest completion time. In addition, the with error experiments are generally slightly longer or comparable to the without error results, with the difference typically within 0.1 s, which is negligible. Moreover, the trend of completion time variation with respect to the frame horizon is consistent with that observed in the without error experiments.

TABLE V. TASK COMPLETION TIMES UNDER DIFFERENT PREDICTION HORIZONS

Observation Type	Horizon (frames)	Completion Time	
Baseline	None	4.9	
Current Only	0	4.6	
Current + prediction		With Error	Without Error
	1	4.7	4.6
	2	4.3	**4.3**
	3	4.5	4.4
	4	4.5	4.4
	5	4.6	4.5

These findings suggest that access to target position information—whether current or predicted—is crucial for effective policy learning. The significant performance improvement observed in the "Current Only" setting indicates that even static spatial cues provide meaningful guidance for the control strategy. Moreover, the "Current + Prediction" setup showed further improvements, and the predicted positions—even with added random error—still contributed to higher success rates.

V. DISCUSSION

This study introduced predicted human facial motion into imitation learning control. Experimental results showed that incorporating predictive observations improved task performance, although the overall improvement remained limited.

In the baseline condition, the robot lacked explicit mouth position information, often moving the cube to non-mouth facial regions with successes occurring mostly by chance. Providing the current mouth position offered clear spatial cues and improved success rates, but frequent failures revealed the absence of higher-level feedback to reliably associate observations with the intended goal. Performance also depended on the prediction horizon: too few frames limited foresight, while too many frames accumulated errors, leading to detours or overly conservative actions, with 2–3 frames appearing to provide the best trade-off, although 50 trials are insufficient for a conclusive evaluation. Overall, despite measurable improvements, the approach remains open-loop imitation learning without explicit goal constraints or dynamic error-correction, and all experiments were conducted in simulation with a non-physical facial model, leaving safety, collision handling, and real-world validation unaddressed.

Future work will focus on improving safety, enhancing goal-oriented control, and moving toward real-world validation. We plan to add physical properties and collision volumes to the simulated face model, enabling the policy to perceive potential collision risks during training and execution, thereby improving safety. In addition, we will explore goal-oriented imitation learning strategies to reduce errors and avoid the "eye-poking" problem, for example by explicitly defining the mouth region as a target and integrating avoidance or safety rewards into the policy. After strengthening safety and collision modeling, we will extend the evaluation from simulation to real-world scenarios to verify the method's feasibility and robustness for practical face-related robotic tasks.

VI. CONCLUSION

This study proposed a Transformer-based approach for predicting human facial movements and integrating these predictions into an imitation learning framework for face-related robotic control tasks. The cross-attention Transformer achieved higher landmark prediction accuracy and better image quality than alternative methods.

In imitation learning experiments, incorporating facial motion information significantly improved task performance. Using facial trajectories extracted from real human demonstrations, the success rate improved from 42% to 60%, and further to 74% with predicted landmarks. The 2-frame prediction horizon produced the fastest task completion times under these realistic motion conditions.

These results demonstrate that integrating predicted facial motion into the observation space can substantially enhance imitation learning performance in dynamic human-robot interaction. Future work will focus on incorporating safety and collision modeling, enhancing goal-oriented imitation learning strategies, and extending evaluation from simulation

to real-world scenarios to improve robustness and practical applicability.

ACKNOWLEDGMENT

This achievement was supported by JST SPRING, Grant Number JPMJSP2124.

REFERENCES

[1] U. Sania and S. H. Naaz, "Advancing Human-Robot Interaction for Assistive Technologies in Healthcare," IHERT, vol. 6, no. 2, pp. 759–766, Dec. 2024.

[2] R. K. Jenamani, T. Silver, B. Dodson, S. Tong, A. Song, Y. Yang, Z. Liu, B. Howe, A. Whitneck, and T. Bhattacharjee, "FEAST: A Flexible Mealtime-Assistance System Towards In-the-Wild Personalization," arXiv preprint arXiv:2506.14968, 2025. [Online]. Available: https://arxiv.org/abs/2506.14968

[3] H. W. Chang et al., "Design of a Breakaway Utensil Attachment for Enhanced Safety in Robot-Assisted Feeding," arXiv preprint arXiv:2502.17774, 2025.

[4] D. Park et al., "Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned," Robotics and Autonomous Systems, vol. 124, pp. 103344, 2020, doi: 10.1016/j.robot.2019.103344.

[5] A. Candeias, T. Rhodes, M. Marques, J. P. ao Costeira, and M. Veloso, "Vision Augmented Robot Feeding," in Proc. Eur. Conf. Comput. Vis. Workshops (ECCV Workshops), Sept. 2018.

[6] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," arXiv preprint arXiv:2304.13705, 2023. [Online]. Available: https://arxiv.org/abs/2304.13705

[7] F. Xie, A. Chowdhury, M. De Paolis Kaluza, L. Zhao, L. Wong, and R. Yu, "Deep imitation learning for bimanual robotic manipulation," Advances in Neural Information Processing Systems, vol. 33, pp. 2327–2337, 2020.

[8] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: a survey of learning methods," ACM Comput. Surveys, vol. 50, no. 2, pp. 1–35, 2017.

[9] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: challenges, methods and progress," Artificial Intelligence, vol. 297, p. 103500, 2021.

[10] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, "Multimodal motion prediction with stacked transformers," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 7577–7586.

[11] S. Tankasala and M. Pryor, "Accelerating trajectory generation for quadrotors using transformers," in Proc. 5th Annu. Conf. Learn. Dyn. Control (L4DC), PMLR vol. 211, 2023, pp. 1–12.

[12] A. Bucker, L. Figueredo, S. Haddadin, A. Kapoor, S. Ma, and R. Bonatti, "Reshaping robot trajectories using natural language commands: A study of multi-modal data alignment using transformers," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2022, pp. 978–984.

[13] L. Zhuang, J. Zhao, Y. Li, Z. Xu, L. Zhao, and H. Liu, "Transformer-Enhanced Motion Planner: Attention-Guided Sampling for State-Specific Decision Making," IEEE Robot. Autom. Lett., vol. 9, no. 10, pp. 8794–8801, Oct. 2024, doi: 10.1109/LRA.2024.3450305.

[14] E. V. Mascaro, S. Ma, H. Ahn, and D. Lee, "Robust Human Motion Forecasting using Transformer-based Model," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2022, doi: 10.1109/IROS47612.2022.9981877.

[15] Z. Wang, "Path planning of intelligent tennis ball picking robot integrating twin network target tracking algorithm," Scientific Reports, vol. 15, Article no. 20668, 2025, doi: 10.1038/s41598-025-04865-w.

[16] S. Yin, S. Wang, G. Peng, et al., "Capturing Spatial and Temporal Patterns for Facial Landmark Tracking through Adversarial Learning," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2019, pp. 1010–1017.

[17] P. Micaelli, A. Vahdat, H. Yin, J. Kautz, and P. Molchanov, "Recurrence Without Recurrence: Stable Video Landmark Detection With Deep Equilibrium Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 22814–22825.

[18] S. Yin, S. Huan, S. Wang, J. Hu, T. Guo, B. Yin, and C. Liu, "1DFormer: a Transformer Architecture Learning 1D Landmark Representations for Facial Landmark Tracking," arXiv preprint arXiv:2311.00241, 2024. [Online]. Available: https://arxiv.org/abs/2311.00241

[19] K. Gao et al., "MuST: Multi-Head Skill Transformer for Long-Horizon Dexterous Manipulation with Skill Progress," arXiv preprint arXiv:2502.02753, 2025. [Online]. Available: https://arxiv.org/abs/2502.02753

[20] J. H. Park, W. Choi, S. Hong, H. Seo, J. Ahn, C. Ha, H. Han, and J. Kwon, "Hierarchical Action Chunking Transformer: Learning Temporal Multimodality from Demonstrations with Fast Imitation Behavior," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2024, pp. 12648–12654, doi: 10.1109/IROS58592.2024.10802845.

[21] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, et al., "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019.

[22] M. Murooka, T. Motoda, and R. Nakajo, "RoboManipBaselines," GitHub repository, ver. 1.0.0, Dec. 2024. [Online]. Available: https://github.com/isri-aist/RoboManipBaselines