Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Automated Synthesis of Facial Mechanisms for Conversational Animatronic Robots

Zongzheng Zhang*[1,2], Zi Lin*[1], Jiawen Yang[1], Ziqiao Peng[1],
Junyan Lao[1], Lin Cheng[4], Huazhe Xu[3], Hang Zhao[3], Hao Zhao†[1,2]

[1] Institute for AI Industry Research (AIR), Tsinghua University [2] Beijing Academy of Artificial Intelligence (BAAI)
[3] Institute for Interdisciplinary Information Sciences(IIIS), Tsinghua University [4] Beihang University
* Equal contribution †Corresponding author
Automated-facial-mechanisms-synthesis.github.io

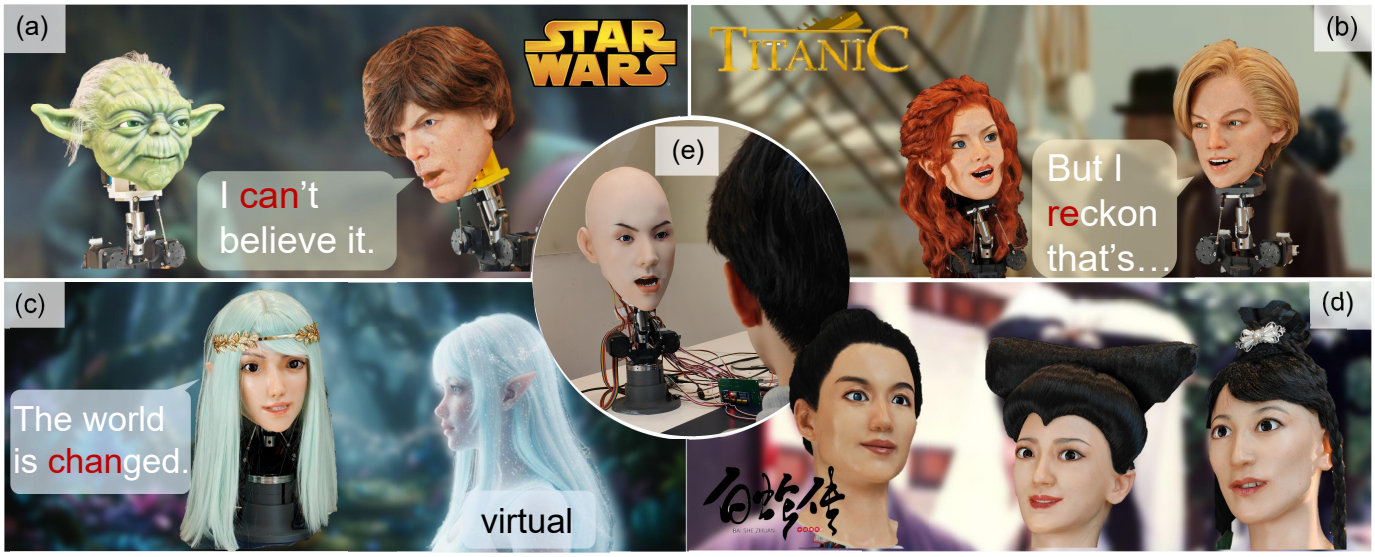

Fig. 1: We demonstrate our end-to-end physical conversational face system across diverse multi-round interactions: (a) reenactments of *Star Wars* (Yoda–Luke Skywalker) and (b) *Titanic* (Rose–Jack); (c) a dialogue between a real elf and a virtual elf; (d) a three-character encounter from the Chinese folktale *Legend of the White Snake*; and (e) human–robot dialogue.

Abstract*—Animatronic faces are a central component of socially interactive robots, enabling rich nonverbal communication through facial articulation. However, state-of-the-art animatronic faces are typically tailored systems: each new facial geometry requires extensive manual mechanical redesign, making large-scale personalization prohibitively slow and costly. In this work, we pursue automated and scalable mechanical face synthesis, aiming to rapidly generate a physically realizable facial mechanism for a wide range of facial geometries. We introduce a parametric, linkage-driven mechanical face template whose topology and actuator layout are explicitly parameterized to support systematic scaling and retargeting across diverse facial morphologies. Building on this template, we propose a hierarchical automatic design algorithm that takes a single 2D portrait as input, reconstructs a target 3D face, and synthesizes a collision-free, manufacturable internal mechanism. The algorithm combines anatomy-guided feasible motion volumes, Action Unit (AU)-derived trajectory-based expressiveness objectives, and a collision-driven outer-loop refinement strategy. Beyond hardware synthesis, we argue that future mechanical faces deployed at scale must engage in bidirectional, multi-turn conversation, rather than functioning solely as speaking or listening heads. To this end, we develop a dual-identity conversational facial motion synthesis framework that jointly models speaking and listening behaviors from audio, producing temporally coherent 3D facial motion suitable for physical execution. We validate our system through extensive experiments, including (i) quantitative evaluation of automatic mechanism synthesis across diverse facial geometries, (ii) comparisons against manual mechanical design, (iii) benchmarks on conversational facial motion synthesis and real-time deployment, and (iv) perceptual user studies.

I. Introduction

Imagine a future where thousands of robots coexist with humans, each endowed with a distinct face, identity, and personality. Rather than sharing a single canonical appearance, these robots would exhibit thousands of unique mechanical faces (from realistic human likenesses to stylized fictional characters) capable of engaging in rich, multi-turn conversations. As illustrated in Fig. 1, such animatronic faces would not merely speak, but participate in bidirectional dialogue, react as listeners, and sustain expressive interactions across diverse narratives and social contexts. Realizing this vision of

massively personalized, conversational mechanical faces represents a foundational step toward socially integrated robotics.

However, despite decades of progress in animatronic face design [51, 52, 10], today's state-of-the-art systems remain fundamentally tailored [41, 117, 111]. High-fidelity platforms achieve impressive realism through carefully engineered actuation layouts, transmission systems, and handcrafted linkages [68, 24, 58]. Yet each of these faces is designed for a single fixed geometry. Adapting such systems to a new facial morphology typically requires extensive manual mechanical redesign, iterative prototyping, and expert intervention. As a result, current approaches scale poorly: they cannot feasibly support the rapid creation of hundreds or thousands of mechanically distinct faces, let alone support the diverse characters demonstrated in Fig. 1. This reliance on custom mechanical design forms a critical bottleneck that prevents animatronic faces from scaling beyond isolated demonstrations.

To overcome this limitation, we argue that animatronic faces must be treated not as handcrafted artifacts, but as automatically synthesizable systems. Our core design principle is simple: instead of redesigning mechanisms from scratch for each face, we start from a parametric mechanical face template whose topology, actuator layout, and kinematic structure are explicitly parameterized. Given this template, we seek an automated pipeline that can rapidly adapt the internal mechanism to any given facial geometry, producing a collision-free, manufacturable design with minimal human intervention. In this paradigm, mechanical face design becomes a computational problem [21, 17, 94, 54].

Building on this principle, we introduce an end-to-end automated mechanical face synthesis pipeline. Starting from a single 2D portrait, our system reconstructs a 3D facial geometry and initializes a modular linkage-driven face template. We then perform hierarchical optimization to adapt the template to the target face. **At the module level**, we synthesize kinematic structures that maximize expressiveness under anatomy-guided feasible motion volumes and AU-derived trajectory objectives. **At the assembly level**, we resolve spatial conflicts using a collision-driven outer-loop refinement strategy that iteratively adjusts base poses or schedules expression amplitudes. This hierarchical design tightly couples kinematic synthesis with engineering validation, enabling the automatic generation of high-DoF facial mechanisms that are both expressive and physically realizable. As demonstrated in Fig. 1, this approach generalizes across human, stylized, and non-human faces, including characters with drastically different morphologies.

Crucially, scalable hardware alone is not sufficient. If animatronic faces are to be deployed at scale in human environments, they must function as conversational agents, not merely as talking heads. Yet most existing facial motion generation and mapping methods are designed for a single face and single-speaker setting [117, 111, 42, 102, 58]. They generate speech-driven expressions without modeling listening behaviors or multi-turn interaction, and are often ill-suited for physical execution. In contrast, we posit that future mechanical faces should participate in dialogue, alternating naturally between speaking and listening roles. To this end, we introduce a dual-identity conversational facial motion synthesis framework that jointly models speaking and listening behaviors from audio, producing temporally coherent 3D facial motion suitable for real-time robotic actuation. We further develop a semantic, region-wise mapping strategy that robustly translates synthesized facial motion into actuator commands, bridging the gap between digital animation and physical execution.

We validate our approach through extensive experiments spanning both hardware and software. **On the mechanical side**, we quantitatively evaluate the success rate, convergence speed, and expressiveness of our automated design pipeline across diverse facial geometries, and compare against manual design and alternative optimization strategies. **On the interaction side**, we benchmark our conversational facial motion synthesis against state-of-the-art talking-head methods in both speaker and listener roles, and demonstrate real-time deployment on physical robots. **Finally**, perceptual user studies confirm that our end-to-end system delivers more natural, engaging, and socially convincing interactions. Together, these results demonstrate that our framework provides a scalable and practical path toward the vision of thousands of personalized, conversational animatronic faces.

II. Related Work

A. Mechanical Face Platform

The evolution of mechanical face design has transitioned from early exploration of actuation modalities to today's hyper-realistic platforms. Early research established the feasibility of FACS-based motion using pneumatic muscles and micro-actuators [51, 52, 53, 36, 37, 5, 46, 78], later incorporating bio-inspired materials and tendon-driven mechanisms to improve compliance [35, 97, 34, 27, 58]. Subsequent efforts focused on system integration, embedding dense DoFs into full-body humanoids [48, 75, 56, 1, 64] and android heads [34, 27, 31], while investigating social engagement through developmental platforms like Cog and Affetto [7, 72, 44, 2]. Recently, the field has shifted toward high-fidelity architectures with refined transmission systems, exemplified by platforms such as Abel, Ameca, Nikola, and EMO [16, 77, 26, 10, 41, 103, 104]. Latest designs further explore diverse points in the design space, from robust hybrid actuation schemes (Morpheus) to low-cost soft solutions [90, 68, 24, 117, 111]. Despite these advancements, a critical limitation persists: state-of-the-art systems are bespoke designs engineered for a single fixed geometry. Adapting these sophisticated architectures to new 3D faces typically necessitates substantial manual redesign, a scalability bottleneck our work aims to resolve.

B. Automatic Robot Design

Computer graphics has extensively explored synthesizing mechanisms from motion specifications, ranging from planar linkages and characters [17, 3, 94] to automata derived from motion capture [9, 115]. Complementary tools have facilitated rapid fabrication [74, 8] and template retargeting [107], though

these methods typically prioritize kinematic feasibility over the internal interference required for physical operation. In robotics, optimization extends to dynamics, with task-based methods tuning cable-driven mechanisms [59, 49, 45, 86]. Furthermore, morphology-control co-optimization utilizing differentiable, RL, and graph-based frameworks [101, 32, 89, 38, 96, 112, 50, 62, 93] has been applied across domains including locomotion [28, 33], soft robotics [66, 40], manipulation [73, 30, 29, 105, 54], and multicopters [21, 100]. However, these strategies often yield coarse topologies, lacking the precision to embed complex linkages into constrained facial volumes. Specific to facial robotics, seminal works [6, 43] address the coupling of rigid actuators and deformable skin. Yet, they treat the kinematic structure as fixed, optimizing only placement. In contrast, our framework unifies kinematic synthesis with engineering validation to automate the design of high-DOF facial mechanisms.

C. Talking Head Generation and Expression Mapping

While 2D audio-driven generation offers visual fidelity [87, 108, 55, 85, 65, 113], its high latency and limited perspective robustness [114, 116, 102, 42, 110] make 3D synthesis preferable for robotic actuation [18, 88, 14, 111]. 3D synthesis has evolved from deterministic lip-sync [25, 99, 63] to emotional expressiveness [81, 19, 67] and cross-identity generalization [70, 23, 11]. Crucially, recent research shifts from monologue [80, 82, 84] to dyadic interaction, employing dual-turn frameworks to model reciprocal dynamics and listening behaviors [83, 76, 95, 15, 12]. To bridge the gap to physical robots, mapping approaches have progressed from geometric correspondence [10, 24, 98, 39, 117] to semantic representations and data-driven optimization [69, 60, 106, 109, 104, 58]. Unlike prior works restricted to single-speaker generation or frame-wise mapping, we strictly enforce a joint modeling of dual-turn synthesis and semantic region-wise mapping to enable robust multi-turn conversational interaction.

III. Automated Scalable Mechanism Design

In this section, we present an end-to-end pipeline for **automated hardware design** of a scalable robotic face. We first introduce a **scalable hardware template** derived from a canonical mean face model, where the actuator layout and linkage topology are parameterized to support rapid adaptation across diverse facial morphologies (Sec. III-A). Building on this template, we then propose a **hierarchical automatic design algorithm** (Algorithm 1) that takes a single 2D portrait as input, reconstructs the target 3D facial geometry, and progressively optimizes the template parameters under manufacturability and non-interference constraints (Sec. III-B).

A. Parametric Face Template Design

We develop a modular, parametric mechanical-face template that serves as the starting point of our automated design pipeline. The face template is composed of four modules: **eyebrow**, **eyes**, **mouth**, and **jaw** (Fig. 2). The eyebrow module contains two kinds of motion: the vertical movement of the

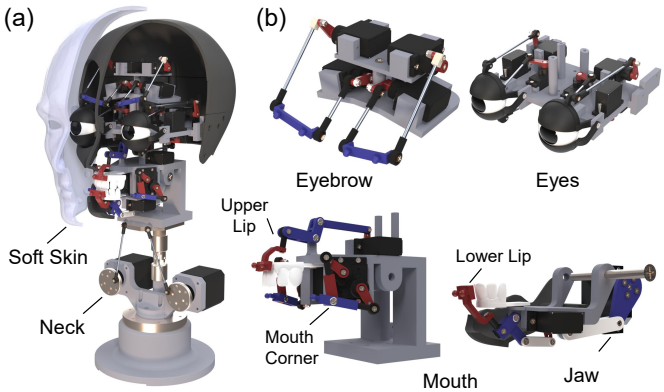

Fig. 2: **Mechanical face template.** (a) Full assembly of the linkage-driven robotic face with a soft skin and a 3-DoF neck module. (b) Exploded view of the four modular facial mechanisms—eyebrow, eyes, mouth, and jaw.

eyebrow ridge and the brow center. Both are realized by a spatial six-bar mechanism with 2 DoF. The eye module contains 8 DoF. It comprises four eyelid linkages (upper/lower lids for both eyes) and two rotational DoFs per eyeball. The mouth module integrates mechanisms for the upper/lower lips and mouth corners. Both the upper and lower lips are actuated by spatial four-bar linkages whose end-effector motions follow vertical trajectories to match the predominant anatomical direction of lip deformation. The mouth corners are driven by a five-bar mechanism that supports three principal motion modes: upward raising, lateral translation, and downward pulling. The jaw module is actuated by a four-bar mechanism. Finally, head pose is produced by a neck module consisting of a 2 DoF parallel mechanism for nodding/tilting and an additional rotational DoF for yaw.

B. Hierarchical Automatic Design

1) Coarse initialization: Given an input RGB portrait image I, we use the framework [118] to construct a normalized 3D facial mesh $\mathcal{M}$ together with a set of semantically labeled 3D facial landmarks $\mathbf{P}$. We then import $\mathcal{M}$ into a CAD environment and apply a scaling factor to bring the head geometry into a physically plausible metric range. These 3D semantic points are then used for the subsequent module-level coarse initialization (Fig. 3(a)).

Using the scaled semantic landmarks, we perform a module-level coarse initialization to obtain the initial base poses $\mathbf{p}^{\text{base}}$. The **eyes** and **jaw** modules are fully determined at this stage. Specifically, the eye geometry is directly computed by periocular landmark subsets, and the jaw module is anchored by the jawline landmarks, yielding a fixed initialization. In contrast, the **eyebrow** and **mouth** (mouth corners and lips) modules require further kinematic synthesis. We initialize their base poses $\mathbf{p}^{\text{base}}_{\text{brow}}$ and $\mathbf{p}^{\text{base}}_{\text{mouth}}$ by scaling the template anthropometric offsets according to the reconstructed head scale: $\mathbf{p}^{\text{base}}_k \leftarrow \mathbf{p}^{\text{base}}_{k,0} + \beta\,\Delta\mathbf{p}^{\text{tmpl}}_k$, where $\Delta\mathbf{p}^{\text{tmpl}}_k$ denotes the corresponding relative offsets in the mean-face template, and β is the global scaling factor applied to $\mathcal{M}$. Their structural parameters are initialized with the template values, which

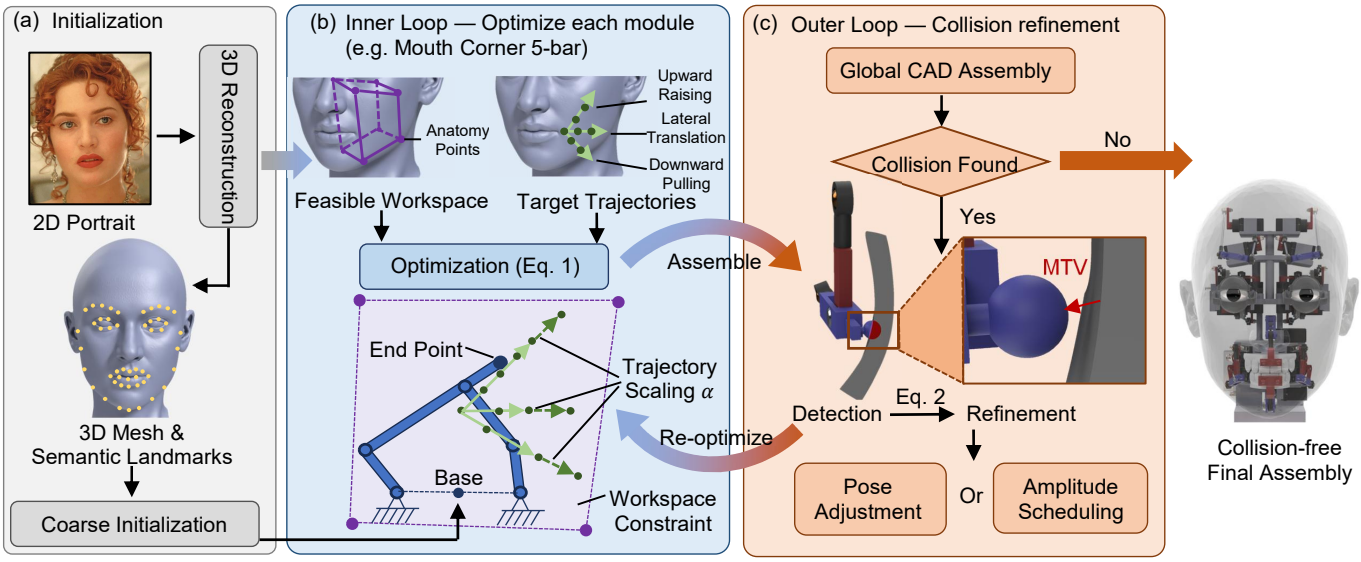

Fig. 3: **Overview of the hierarchical automatic design pipeline.** (a) **Initialization**: From a 2D portrait, we reconstruct a 3D head mesh, semantic landmarks, and initialize module base poses. (b) **Inner loop**: We perform module-wise kinematic synthesis under anatomy-guided feasible volumes and AU-derived trajectories to maximize expressiveness (illustrated with the mouth-corner five-bar mechanism). (c) **Outer loop**: We assemble the global CAD model, detect interferences, and apply MTV-guided base-pose updates (or, if necessary, reduce the amplitude limit) until a collision-free assembly is obtained.

serve as the starting point for the module-wise kinematic design optimization in the next step (see App. D1 for details).

2) Module-wise kinematic design optimization: With the rigid modules (eyes and jaw) geometrically determined in the coarse initialization phase, this step focuses on the kinematic synthesis of the eyebrow and mouth modules. The core challenge lies in synthesizing mechanism parameters that maximize expressive amplitude while strictly adhering to the anatomical spatial constraints imposed by the facial skin.

Anatomy-guided feasible motion volumes. To ensure the synthesized mechanisms operate within bio-plausible limits, we define the feasible workspace Ω_{feas} for the eyebrow, mouth corner, and lip regions (Fig. 3(b); details in App. D2), following the evidence from [20, 13]. We construct Ω_{feas} as the maximal anatomically valid envelope, so that the optimizer retains sufficient freedom to maximize expressiveness. Simultaneously, strictly confining the solution to this valid manifold acts as a regularization term. This accelerates convergence by pruning the search space of physically impossible poses and mitigates the risk of the solver getting trapped in invalid local minima, thereby guiding the optimization toward realizable solutions.

Trajectory-based expressiveness objectives. We derive module-specific motion trajectories from facial Action Units (AUs) [22] to ensure that the designed mechanisms follow anatomically consistent patterns of human facial motion. Concretely, we define a set of canonical target trajectories $\mathcal{B} = \{\mathbf{T}_1, \ldots, \mathbf{T}_K\}$, where each $\mathbf{T}_k$ is represented by a sequence of M sampled points in the local workspace (details in App. D3). We reformulate the goal of maximizing expressiveness as a trajectory scaling problem. For each trajectory $\mathbf{T}_k$, we introduce a scalar amplitude coefficient α_k. Crucially, this scaling is directional: rather than isotropically enlarging the mechanism, it linearly extrapolates the waypoints along the trajectory path $\mathbf{T}_k$ (Fig. 3(b)). By optimizing for larger $\{\alpha_k\}_{k=1}^K$ values, we aim to synthesize mechanisms capable of executing high-amplitude expressions while strictly preserving the semantic direction of the original human movements.

Optimization formulation. For each mechanical structure k, we perform kinematic design optimization in its local coordinate frame. Specifically, using the module base pose $\mathbf{p}_{base}$ from coarse initialization, we transform $\mathbf{T}_k$ to each local frame. We formulate the module-wise mechanism synthesis as a constrained non-linear optimization problem. Given the target trajectories $\mathbf{T}_k$, we seek to determine the optimal structural parameters $\mathbf{\Phi}_k$ (link lengths), the set of discrete joint configurations $\mathbf{\Theta}_k$, and the expression amplitude coefficients α_k. We define a unified cost function $\mathcal{L}$ to balance expressive range, tracking accuracy, and kinematic manipulability:

$$\begin{aligned} \min_{\mathbf{\Phi},\mathbf{\Theta},\boldsymbol{\alpha}} \quad & \mathcal{L} = \omega_{\text{amp}}\mathcal{L}_{\text{amp}} + \omega_{\text{fit}}\mathcal{L}_{\text{fit}} - \omega_{\text{man}}\mathcal{L}_{\text{man}} \\ \text{s.t.} \quad & \mathbf{P}(\boldsymbol{\theta}_{k,m}) \subset \Omega_{\text{feas}}, \quad \forall m \in \{1,\ldots,M\}, \\ & 0 < \alpha_k \le \alpha_k^{\text{limit}}, \quad \forall k \in \{1,\ldots,K\}, \end{aligned} \tag{1}$$

where ω_{amp}, ω_{fit}, and ω_{man} are user-defined weighting factors. The individual objective terms are defined as follows: $\mathcal{L}_{\text{amp}} = 1/\alpha_k$; $\mathcal{L}_{\text{fit}} = \sum_{m=1}^{M} \|f(\boldsymbol{\theta}_{k,m}, \mathbf{\Phi}) - \mathbf{p}_{k,m}(\alpha_k)\|_2^2$, where $\mathbf{p}_{k,m} \in \mathbf{T}_k$; $\mathcal{L}_{\text{man}} = \sum_{m=1}^{M} \sqrt{\det(\mathbf{J}(\boldsymbol{\theta}_{k,m})\mathbf{J}(\boldsymbol{\theta}_{k,m})^\top)}$. Here, m denotes the waypoint index along the trajectory. $\boldsymbol{\theta}_{k,m}$ denotes the instantaneous joint configuration at step m, $f(\cdot)$ represents the forward kinematics, and $\mathbf{J}$ is the geometric Jacobian matrix utilized to maximize workspace manipulability (see App. C for explicit derivations).

3) Global CAD assembly and collision-driven refinement: This step integrates all modules into a global CAD assembly

Algorithm 1 Hierarchical Automatic Design

Require: 2D portrait image I; template face design
Ensure: Final assembly $\mathcal{A}$

$(\mathcal{M}, \mathbf{P}) \leftarrow$ RECONSTRUCT3D(I)
$\{\mathbf{p}^{\text{base}}\} \leftarrow$ COARSEINITIALIZATION($\mathcal{M}, \mathbf{P}$) ▷ Initialization
for iteration $s = 1$ to S **do**
  for mechanism $k \in \{\text{brow}, \text{mouth}\}$ **do** ▷ Inner loop
    $\{\mathbf{T}_k\} \leftarrow$ GLOBALTOLOCAL($\{\mathbf{T}_k\}, \mathbf{p}_k^{\text{base}}$)
    $(\mathbf{\Phi}_k, \mathbf{\Theta}_k, \boldsymbol{\alpha}_k) \leftarrow$ SOLVEEQ. 1($\Omega_{\text{feas}}, \{\mathbf{T}_k\}, \{\alpha_k^{\text{limit}}\}$)
  end for
  $\alpha_k^{\text{limit}} \leftarrow \min(\alpha_k^{\text{limit}}, \alpha_k^*),\ \forall k$ ▷ Outer loop
  $\mathcal{A} \leftarrow$ ASSEMBLECAD($\{\mathbf{p}_k^{\text{base}}\}, \{\mathbf{\Phi}_k\}$)
  $\mathcal{C} \leftarrow$ COLLISIONDETECTION($\mathcal{A}, \{\mathbf{\Theta}_k\}$)
  if $\mathcal{C} = \emptyset$ **then**
    break
  end if
  for each mechanism k with constraints $\mathcal{C}_k$ **do**
    $(\{\mathbf{n}_j, \delta_j\}_{j \in \mathcal{C}_k}) \leftarrow \text{MTV}(\mathcal{C}_k)$
    $\Delta\mathbf{p}_k \leftarrow$ SOLVEEQ. 2($\{\mathbf{n}_j, \delta_j\}_{j \in \mathcal{C}_k}, \epsilon$)
    if $\|\Delta\mathbf{p}_k\|_2 > D_{\max}$ **then** ▷ Amplitude scheduling
      $\alpha_k^{\text{limit}} \leftarrow \gamma\, \alpha_k^{\text{limit}}$
      continue ▷ Skip update, re-optimization next iter
    end if
    $\mathbf{p}_k^{\text{base}} \leftarrow \mathbf{p}_k^{\text{base}} + \Delta\mathbf{p}_k$ ▷ Pose adjustment
  end for
end for
return $\mathcal{A}$

via API and resolves spatial conflicts through a collision-driven outer-loop refinement strategy. For a mechanism k, we sample discrete configurations $\{\boldsymbol{\theta}_{k,m}\}_{m=1}^{M}$ along its optimized trajectory and utilize a collision detection library FCL [79] to identify interferences between the k-th module and the environment (skull and adjacent mechanisms). For every active collision constraint $j \in \mathcal{C}_k$, we compute a **Minimum Translation Vector** $\mathbf{v}_{\text{int}} = \delta_j \cdot \mathbf{n}_j$, where δ_j is the penetration depth and $\mathbf{n}_j$ is the separation normal (Fig. 3(c)). To resolve multiple simultaneous collisions while minimally perturbing the design, we solve a QP problem:

$$\min_{\Delta\mathbf{p}_k} \quad \|\Delta\mathbf{p}_k\|_2^2 \quad \text{s.t.} \qquad \mathbf{n}_j^\top \Delta\mathbf{p}_k \geq \delta_j + \epsilon, \quad \forall j \in \mathcal{C}_k, \tag{2}$$

where ϵ is a safety clearance margin. We accept the pose update only if $\Delta\mathbf{p}_k$ falls within a threshold $D_{\max}$; otherwise, we contract α_k^{limit} via a decay factor $\gamma \in (0, 1)$. The updated $\mathbf{p}_k^{\text{base}}$ is fed back to the inner-loop optimization phase (Sec. III-B2), where $\mathbf{\Phi}_k$ and $\mathbf{\Theta}_k$ are re-optimized in the new local frame. The outer loop iterates until all collisions are satisfied ($\mathcal{C} = \emptyset$), outputting the final assembly $\mathcal{A}$.

IV. Interaction Synthesis and Control

Following the automated mechanical-face adaptation described above, we present the algorithms that **drive interactive expressions on the physical robot**. We first generate **3D facial motion** for multi-round dialogue, modeling both speaking and listening behaviors under conversational context (Sec. IV-A). We then **map** the synthesized digital facial motion to robot-actuable commands (Sec. IV-B). Together, these modules bridge conversational interaction, 3D facial animation, and embodied execution on the mechanical face.

A. Interactive Talking Head Synthesis

Unified dual-speaker interaction framework. The framework (Fig. 4(a)) takes the audio from Speaker-A and Speaker-B as input and directly outputs a temporally coherent sequence of blendshape parameters that drives the synthesized facial motion. First, we use a pre-trained audio encoder Wav2Vec 2.0 [4] to map $\mathbf{A}_A \in \mathbb{R}^{T \times F}$ and $\mathbf{A}_B \in \mathbb{R}^{T \times F}$ into a shared latent space, where T is the sequence length and F is the sampling rate. We then apply a learnable linear projection $\mathbf{W}_a$ to align both streams to a shared feature dimension d:

$$\mathbf{Z}_A = \mathbf{W}_a\big(E_{\text{aud}}(\mathbf{A}_A)\big), \qquad \mathbf{Z}_B = \mathbf{W}_a\big(E_{\text{aud}}(\mathbf{A}_B)\big), \tag{3}$$

where $\mathbf{Z}_A, \mathbf{Z}_B \in \mathbb{R}^{L \times d}$ serve as the unified audio feature. We add temporal positional encoding and style embeddings to get $\tilde{\mathbf{Z}}_A, \tilde{\mathbf{Z}}_B$. Since the two speakers are largely non-overlapping in speech, we regulate the contribution of the two audio streams via a turn-aware gate derived from speech activity. Let $g_A(t) \in [0, 1]$ indicate whether Speaker-A is speaking at frame t. We form a turn-conditioned audio memory for Speaker-A as: $\mathbf{M}_A(t) = g_A(t)\,\tilde{\mathbf{Z}}_A(t) + \big(1 - g_A(t)\big)\,\tilde{\mathbf{Z}}_B(t)$. Then we adopt a Transformer backbone. Specifically, an encoder models long-range dependencies of the turn-conditioned memory. To further stabilize speaking–listening transitions, we apply a turn-aware alignment attention. We decode the full motion using T learnable motion queries with positional encodings. Finally, a regression head predicts Speaker-A blendshape parameters $\hat{\mathbf{Y}} \in \mathbb{R}^{T \times 55}$ (details in App. F).

Dataset construction. While recent conversational interaction datasets [83, 15] utilize FLAME [61] parameters, its expression coefficients are not aligned with semantically interpretable Blendshape [57] controls, which complicates the mapping process. We therefore build a dataset of paired two-speaker audio and per-frame blendshape coefficients.

We collect raw conversation videos from YouTube, and RealTalk [47]. Each video is first segmented into temporally coherent clips using TransNetV2 [92]. For each clip, we separate the two-speaker audio streams with SAM Audio [91], obtaining synchronized audio $\mathbf{A}_A, \mathbf{A}_B$. On the visual side, we run MediaPipe [71] to extract per-frame Blendshape parameters $\mathbf{b} \in \mathbb{R}^{52}$ and head rotation $\mathbf{R} \in \mathbb{R}^3$. The final dataset consists of synchronized tuples $\mathcal{D} = \{(\mathbf{A}_A, \mathbf{A}_B, \mathbf{B}_A, \mathbf{B}_B, \mathbf{R}_A, \mathbf{R}_B)\}$.

B. Human-to-Robot Facial Expression Mapping

Mapping network. At each frame t, the synthesis network predicts $\hat{\mathbf{y}}_t = [\hat{\mathbf{b}}_t; \hat{\mathbf{r}}_t]$. The head motion is realized by the neck module; given $\hat{\mathbf{r}}_t$, we compute the corresponding motor targets $\mathbf{u}_t^{\text{head}}$ via an inverse-kinematics (IK) solver. For the facial expressions, we exploit the semantic structure of the blendshape basis by partitioning the coefficient vector into region-specific subsets, denoted as $\hat{\mathbf{b}}_t = \{\hat{\mathbf{b}}_t^{(k)}\}_{k \in \mathcal{K}}$ with $\mathcal{K} = \{\text{eyebrow}, \text{eyes}, \text{mouth}, \text{jaw}\}$. For each region k, we train a lightweight inverse-mapping MLP that maps the local blendshape coefficients to the corresponding actuator command vector: $\mathbf{u}_t^{(k)} = \pi_k\big(\hat{\mathbf{b}}_t^{(k)}\big)$. Each regressor is trained with a standard Mean Squared Error objective over paired

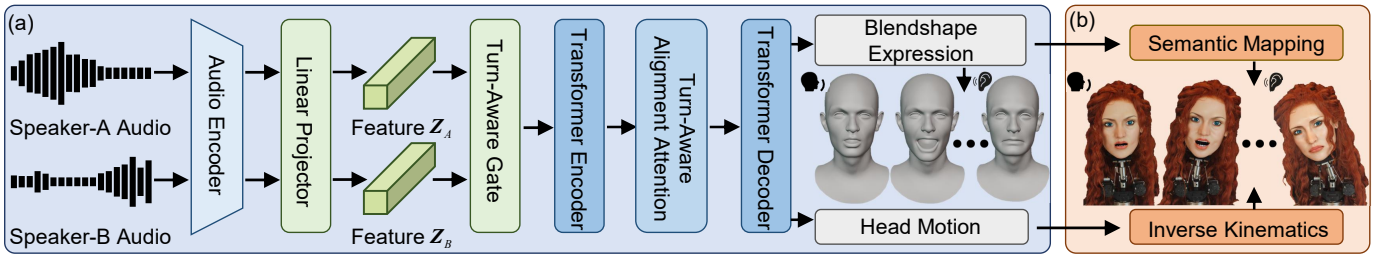

Fig. 4: **Overview of interaction synthesis and control.** (a) Given dual-speaker audio, the talking-head model predicts multi-round facial motion for both speaking and listening. (b) The predicted coefficients are mapped to robot motor commands via region-wise regressors and neck IK, enabling physically feasible multi-round expressions on the mechanical face.

TABLE I: Comparison of optimization strategies for **automatic mechanism design**. We report success rate (SR), convergence time (CT), expressiveness score (Exp $= \sum_k \alpha_k$), and intersection volume (IV). CT and Exp are reported **only** for successful runs; IV is 0 for successful runs and otherwise computed from the final design at the maximum iteration.

Method	SR (%) ↑	CT (s) ↓	Exp ↑	IV (mm^3) ↓
Ours w/o outer loop	20.0	25.7	8.69	2050.7
Global Joint-Opt	0.0	–	–	9690.5
Local + Heuristic	33.3	1098.2	8.57	1690.6
Ours	**66.7**	591.3	8.53	**460.4**

data $\{(\mathbf{b}_t^{(k)}, \mathbf{u}_t^{(k)})\}$: $\mathcal{L}_{\text{map}}^{(k)} = \left\| \pi_k(\mathbf{b}_t^{(k)}) - \mathbf{u}_t^{(k)} \right\|_2^2$. Then the final motor command $\mathbf{u}_t$ at frame t is obtained (Fig. 4(b)).

Data collection. For each physical head, we first calibrate the valid actuation ranges. We then generate 3000 motor command vectors using Latin hypercube sampling within these bounds to ensure broad and uniform coverage of the feasible actuation space. For each sampled command $\mathbf{u}_i$, we capture a synchronized image I_i of the face. From each image I_i, we estimate the corresponding facial parameters $(\mathbf{b}_i, \mathbf{r}_i)$ using MediaPipe [71] and obtain the final supervision tuples $\{(\mathbf{u}_i, \mathbf{b}_i, \mathbf{r}_i)\}$ for training the inverse mapping networks.

V. Experiment

A. Experimental Setup

Hardware platform. Each robotic head provides 21 independently actuated facial degrees of freedom, driven by GuoHua 9g digital micro servos with a 180° operating range. To support higher-torque head orientation, the neck is actuated by three NEMA 17 stepper motors. All structural components were fabricated using Polylactic Acid (PLA) via FDM 3D printing. The silicone skin is magnetically attached to the underlying frame, enabling rapid installation and replacement.

Onboard deployment. All inference runs on an NVIDIA Jetson AGX Orin (32 GB). For actuation, the Orin communicates via the I^2C protocol with cascaded PCA9685 PWM drivers, updating servo commands at 30 Hz. The Orin also interfaces with an eye-mounted camera, a microphone for audio capture, and an onboard speaker for audio playback.

B. Evaluation of the Automatic Design Flow

Hierarchical optimization efficacy. We evaluate our hierarchical optimization framework on 15 heads with diverse facial geometries. We compare against three baselines: (i) *Ours w/o outer loop* (local-only optimization), which performs module-level optimization but disables the outer-loop assembly refinement; (ii) *Global Joint-Opt*, which jointly optimizes all design variables using a single objective that combines position-keeping and collision-penalty terms; and (iii) *Local + Heuristic Repulsion*, which augments local optimization with a hand-crafted repulsion vector to resolve collisions during assembly (App. D5). All methods are optimized using L-BFGS. Tab. I summarizes the mean performance over 15 design instances, where all methods share the same initialization. *Local-only* converges rapidly (25.7s) but fails in compact geometries, validating the indispensability of assembly-level refinement. The *Global Joint-Opt* baseline fails in all cases: the coupled objective is highly non-convex, and collisions lack informative gradients for complex assemblies, leading to poor convergence. While *Local + Heuristic* improves feasibility (5/15), its reliance on suboptimal centroid-based repulsion significantly prolongs convergence (1098.2s) and fails to resolve complex interlocks requiring amplitude scheduling. In contrast, our hierarchical framework achieves the highest success rate (10/15) and converges 1.9× faster than the heuristic baseline. Notably, the runtime difference between *Local-only* and *Ours* (25.7s vs. 591.3s) indicates that collision resolution dominates the computational cost; however, this investment is strictly justified by the substantial gain in realizability while maintaining comparable expressiveness.

Fig. 5(a) shows qualitative results across diverse face geometries. We highlight two heads that deviate substantially from the template in both morphology and scale. **Left:** *Jack* exhibits a sharper, more pointed chin than the template, which leaves less internal clearance and imposes stricter placement constraints for the mouth-corner and lower-lip mechanisms. **Right:** *Yoda* has a markedly non-human facial structure; nevertheless, our pipeline successfully synthesizes a feasible internal mechanism layout. These examples demonstrate that our automatic design flow can handle large geometric variations and generalizes beyond human-like faces, indicating broad applicability and robustness.

Comparative efficiency against manual design. We further compare our automatic design flow with expert manual design on a compact head geometry (scaling factor $\beta = 0.84$ relative to the template), which leaves limited internal volume for mechanisms. Two senior mechanical designers independently

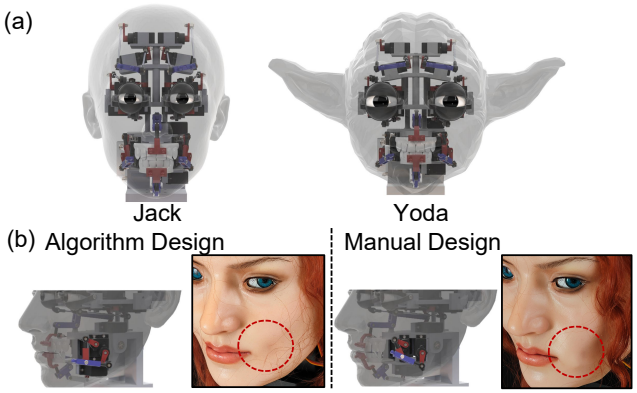

Fig. 5: (a) **Algorithm design versatility**: CAD Assemblies for *Jack* and *Yoda*. (b) **Algorithm vs. Manual**: Comparison of mouth-corner mechanisms (red circles) on a compact head.

TABLE II: Comparison under **speaker** and **listener** modes on our dataset. "RT" means whether real-time use is supported.

Method	Speaker				Listener		RT
	LVE ↓	MSE ↓	PDD ↓	SID ↑	FDD ↓	PDD ↓	
FaceFormer	4.25	0.69	7.32	0.71	–	–	✓
EmoTalk	**2.98**	0.68	6.88	3.03	–	–	✓
SelfTalk	5.34	0.73	7.03	2.98	–	–	✓
DualTalk	3.79	0.57	8.74	**3.36**	29.77	9.45	✗
Ours	3.04	**0.38**	**5.63**	3.29	**21.12**	**5.81**	✓

performed manual layouts and iterative adjustments, including rework triggered by interferences discovered during assembly. The average manual design time is 22.8 h, substantially longer than our optimization runtime (11.7 min). Fig. 5(b) compares the resulting mouth-corner mechanism behaviors. Manual designs, driven by trial-and-error and implicit design heuristics, achieve weaker upward mouth-corner expressiveness and less skin-conforming motion trajectories, leading to reduced visual fidelity. In contrast, our method explicitly optimizes expressiveness under geometric constraints and produces a tighter, more consistent upward expressiveness.

C. Evaluation of Conversational Motion Synthesis

Baselines and metrics. We compare against three **speaker-only** methods (FaceFormer [25], EmoTalk [81], SelfTalk [80]) and a speaker-listener baseline [83]. All methods are trained on our dataset. For the speaker, we evaluate lip synchronization (LVE), overall expression accuracy (MSE), head-pose dynamics (PDD), and motion diversity (SID). For the listener, we evaluate upper-face motion dynamics (FDD) and PDD.

Quantitative results. As shown in Tab. II, our framework demonstrates superior performance across both conversational roles. In speaker mode, we achieve the lowest MSE and PDD, indicating that our generated expressions are both geometrically accurate and temporally stable while maintaining lip-sync quality. In listener mode, we outperform the dyadic baseline [83]. Crucially, [83] relies on the partner's future motion, restricting it to offline processing. In contrast, our model infers reactive behaviors solely from the two speakers' audio, thereby enabling real-time inference.

TABLE III: Comparison for **mapping** grouped by facial representations. FPS is measured on Jetson AGX Orin.

Metric	Landmarks			FLAME	Blendshapes	
	NNR	Norm. [10] + MLP	Norm. [41] + MLP	MLP	MLP	Ours
LSE-D ↓	15.59	14.52	13.74	13.06	12.68	**10.61**
LSE-C ↑	0.34	1.94	2.39	2.42	2.58	**3.19**
FPS ↑	6.36	6.31	6.19	2408.55	**2478.26**	2396.31

D. Evaluation of Expression Mapping Results

Baselines and metrics. We compare our semantic region-wise mapping against baselines spanning three common facial representations. **Landmarks**: (i) nearest-neighbor retrieval that matches each avatar landmark to the closest sample in the robot dataset, and two learning-based variants that reduce the avatar-robot gap via (ii) [10] normalization and [41] normalization followed by an MLP regressor. **FLAME**: an MLP that maps per-frame FLAME parameters to motor commands. **Blendshapes**: per-frame blendshape parameters to motor commands. All learning-based methods are trained with an MSE loss. We assess the robot's audio–lip synchronization using Lip Sync Error Distance (LSE-D) and Lip Sync Error Confidence (LSE-C) [87], as well as on-device FPS. Full-face expression fidelity is evaluated via our user study (Sec. V-F).

Quantitative results. Tab. III shows comparisons across the baselines. Our semantic region-wise mapping achieves the best audio–lip synchronization while maintaining high throughput sufficient for real-time conversational interaction. Normalization-based landmark methods [111, 10, 41] reduce the sim-to-real gap, but their performance is highly sensitive to the robot's facial geometry and the camera viewpoint. In addition, landmark pipelines introduce extra inference latency because they require rendering a 2D avatar frame and running landmark detection before generating motor commands. In contrast, FLAME and blendshape parameters provide identity-agnostic representations. However, a single per-frame MLP underperforms our approach because it must learn cross-region correspondences implicitly from data. By encoding semantic correspondences as an explicit prior, our network achieves superior synchronization with a smaller model.

E. End-to-End Conversation Demonstration

We show our end-to-end system in two interaction scenarios (Fig. 6). **Human-robot interaction.** We integrate the Doubao Voice Model as the conversational engine, enabling fluid, low-latency spoken dialogue with human users (Fig. 6(a)). **Dyadic robot-robot dialogue.** We stage a hypothetical reunion scene from *Titanic*. Driven by input audio, the system generates high-fidelity lip synchronization for the speaker (*Rose*) while synthesizing contextually congruent reactive expressions for the listener (*Jack*) (Fig. 6(b)).

End-to-end build time breakdown. Tab. IV details the mean time required to realize a new character, from mechanism synthesis to a fully deployable robot. Our framework significantly compresses the mechanism design and mapping (Stages 1, 5, 6), which have traditionally served as the primary temporal bottlenecks in robotic customization. While the physical realization (Stages 2, 3, 4) currently occupies the majority

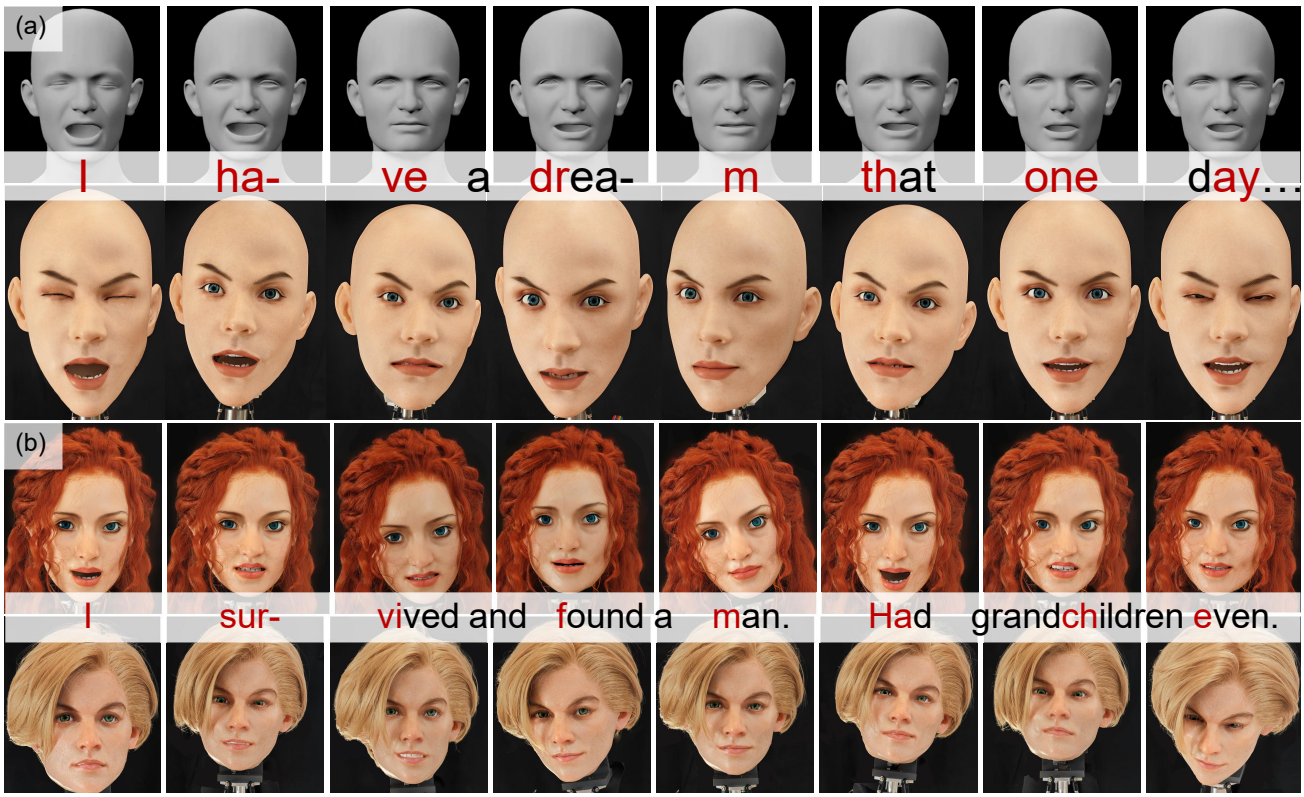

Fig. 6: **End-to-end system visualization across interaction scenarios.** (a) **Human-Robot Interaction**: Real-time conversational gestures during a user-robot dialogue. The top row is the expression synthesis. (b) **Dyadic Role-Play (*Titanic*)**: Reenacting the scene dynamics with expressions corresponding to the dialogue between *Rose* (speaker) and *Jack* (listener).

TABLE IV: **End-to-end time breakdown.** 1: automatic design optimization; 2: 3D printing; 3: silicone face skin; 4: assembly; 5: mapping dataset collection; 6: mapping network training. The *total* time assumes parallel execution of Stages 2 and 3.

Stage	1	2	3	4	5	6	Total
Time (h)	0.16	21.47	20.50	3.83	0.25	0.13	25.84

of the timeline (measured with a single operator and machine), these processes are inherently parallelizable and can be further accelerated through scaled manufacturing resources.

F. User study

To evaluate the perceived quality of our real-time interaction pipeline, we conducted a perceptual study using four representative dialogue excerpts. We recruited 100 participants. For each excerpt, participants were shown the reference clip together with five robot interaction variants: *Ours*, *speaker-only*, *no neck motion*, *random neck motion*, and *mouth-only mapping*. Participants then rated the **overall performance** on a 10-point scale. Fig. 7(a) shows that *Ours* achieves the highest overall scores. The ablations suggest three consistent findings: (i) listener backchannel cues and subtle facial responses substantially improve perceived interaction quality, (ii) coordinated neck motion makes the delivery more natural than facial motion alone [111], and (iii) realism cannot be explained by lip synchronization alone [42, 102], as full-face expressions provide essential cues beyond the mouth region.

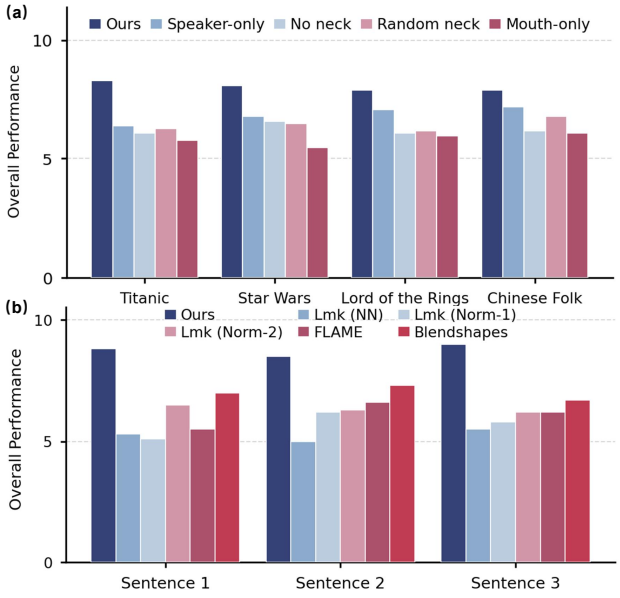

Fig. 7: **User study results.** (a) Overall interaction performance on four excerpts. (b) Mapping quality on three sentences.

We also evaluated mapping quality using three sentences. Fig. 7(b) shows that our mapping produces more realistic and natural expressions, with a smaller gap to the reference motion(details in App. I).

VI. Conclusion

We present a unified framework that enables scalable synthesis for animatronic personalization. By coupling a collision-driven automated hardware design pipeline with a real-time conversational motion engine, we achieved rapid 2D-to-3D physical realization and immersive bidirectional interaction. This work lays the critical groundwork for a future where thousands of robots with distinct identities can be efficiently manufactured and integrated into human society.

References

[1] Ho Seok Ahn, Dong-Wook Lee, Dongwoon Choi, Duk-Yeon Lee, Manhong Hur, and Hogil Lee. Designing of android head system by applying facial muscle mechanism of humans. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pages 799–804. IEEE, 2012.

[2] Brian Allison, Goldie Nejat, and Emmeline Kao. The Design of an Expressive Humanlike Socially Assistive Robot. *Journal of Mechanisms and Robotics*, 1(1), 2008.

[3] Moritz Bächer, Stelian Coros, and Bernhard Thomaszewski. Linkedit: Interactive linkage editing using symbolic kinematics. *ACM Transactions on Graphics (TOG)*, 34(4):1–8, 2015.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[5] Karsten Berns and Jochen Hirth. Control of facial expressions of the humanoid robot head ROMAN. In *2006 IEEE/RSJ international conference on intelligent robots and systems*, pages 3119–3124. IEEE, 2006.

[6] Bernd Bickel, Peter Kaufmann, Mélina Skouras, Bernhard Thomaszewski, Derek Bradley, Thabo Beeler, Phil Jackson, Steve Marschner, Wojciech Matusik, and Markus Gross. Physical face cloning. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012.

[7] Rodney A Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M Williamson. The Cog project: Building a humanoid robot. In *International workshop on computation for metaphors, analogy, and agents*, pages 52–87. Springer, 1998.

[8] Jacques Calì, Dan A Calian, Cristina Amati, Rebecca Kleinberger, Anthony Steed, Jan Kautz, and Tim Weyrich. 3D-printing of non-assembly, articulated models. *ACM Transactions on Graphics (TOG)*, 31(6):1–8, 2012.

[9] Duygu Ceylan, Wilmot Li, Niloy J Mitra, Maneesh Agrawala, and Mark Pauly. Designing and fabricating mechanical automata from mocap sequences. *ACM Transactions on Graphics (TOG)*, 32(6):1–11, 2013.

[10] Boyuan Chen, Yuhang Hu, Lianfeng Li, Sara Cummings, and Hod Lipson. Smile like you mean it: Driving animatronic robotic face with learned models. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2739–2746. IEEE, 2021.

[11] Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Yuan Zhang, Pengfei Wan, Di Zhang, and Shuai Li. Cafe-talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control. *arXiv preprint arXiv:2503.14517*, 2025.

[12] Junjie Chen, Fei Wang, Zhihao Huang, Qing Zhou, Kun Li, Dan Guo, Linfeng Zhang, and Xun Yang. Towards Seamless Interaction: Causal Turn-Level Modeling of Interactive 3D Conversational Head Dynamics. *arXiv preprint arXiv:2512.15340*, 2025.

[13] Yuming Chong, Jingyu Li, Xinyu Liu, Xiaojun Wang, Jiuzuo Huang, Nanze Yu, and Xiao Long. Three-dimensional anthropometric analysis of eyelid aging among Chinese women. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 74(1):135–142, 2021.

[14] Xuangeng Chu, Nabarun Goswami, Ziteng Cui, Hanqin Wang, and Tatsuya Harada. ARTalk: Speech-Driven 3D Head Animation via Autoregressive Model. *arXiv preprint arXiv:2502.20323*, 2025.

[15] Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. UniLS: End-to-End Audio-Driven Avatars for Unified Listening and Speaking. *arXiv preprint arXiv:2512.09327*, 2025.

[16] Lorenzo Cominelli, Gustav Hoegen, and Danilo De Rossi. Abel: integrating humanoid body, emotions, and time perception to investigate social interaction and human cognition. *Applied Sciences*, 11(3):1070, 2021.

[17] Stelian Coros, Bernhard Thomaszewski, Gioacchino Noris, Shinjiro Sueda, Moira Forberg, Robert W Sumner, Wojciech Matusik, and Bernd Bickel. Computational design of mechanical characters. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013.

[18] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10101–10111, 2019.

[19] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–13, 2023.

[20] Richard L. Drake, A. Wayne Vogl, and Adam W. M. Mitchell. *Gray's Anatomy for Students*. Churchill Livingstone, Philadelphia, PA, 3rd edition, 2015. ISBN 978-0702051319.

[21] Tao Du, Adriana Schulz, Bo Zhu, Bernd Bickel, and Wojciech Matusik. Computational multicopter design. *ACM Transactions on Graphics (TOG)*, 35(6):1–10, 2016.

[22] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[23] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pages 204–221. Springer, 2024.

[24] Xuanhe Fan, Huijuan Zhao, Shuangjiang He, Li Li, and Li Yu. A Soft-Skin Facial Robot Capable of Real-Time Emotion-Driven Actuation Through Visual Perception. In *International Conference on Intelligent Robotics and Applications*, pages 54–66. Springer, 2025.

[25] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18780, 2022.

[26] Zanwar Faraj, Mert Selamet, Carlos Morales, Patricio Torres, Maimuna Hossain, Boyuan Chen, and Hod Lipson. Facially expressive humanoid robotic face. *HardwareX*, 9:e00117, 2021.

[27] Leopoldina Fortunati, Alessandra Sorrentino, Laura Fiorini, and Filippo Cavallo. The rise of the roboid. *International Journal of Social Robotics*, 13(6):1457–1471, 2021.

[28] Moritz Geilinger, Roi Poranne, Ruta Desai, Bernhard Thomaszewski, and Stelian Coros. Skaterbots: Optimization-based design and motion synthesis for robotic creatures with legs and wheels. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

[29] Kieran Gilday, Dohyeon Pyeon, S Dhanush, Kyu-Jin Cho, and Josie Hughes. Exploiting passive behaviours for diverse musical playing using the parametric hand. *Frontiers in Robotics and AI*, 11:1463744, 2024.

[30] Kieran Gilday, Chapa Sirithunge, Fumiya Iida, and Josie Hughes. Embodied manipulation with past and future morphologies through an open parametric hand design. *Science Robotics*, 10(102):eads6437, 2025.

[31] Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. Erica: The erato intelligent conversational android. In *2016 25th IEEE International symposium on robot and human interactive communication (RO-MAN)*, pages 22–29. IEEE, 2016.

[32] David Ha. Reinforcement learning for improving agent design. *Artificial life*, 25(4):352–365, 2019.

[33] Sehoon Ha, Stelian Coros, Alexander Alspach, Joohyung Kim, and Katsu Yamane. Task-based limb optimization for legged robots. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2062–2068. IEEE, 2016.

[34] Ahsan Habib, Sumit K Das, Ioana-Corina Bogdan, David Hanson, and Dan O Popa. Learning human-like facial expressions for android Phillip K. Dick. In *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 1159–1165. IEEE, 2014.

[35] David Hanson, Giovanni Pioggia, S Dinelli, Fabio Di Francesco, R Francesconi, and Danilo De Rossi. Identity Emulation (IE): Bio-inspired Facial Expression Interfaces for Emotive Robots. In *AAAI Mobile Robot Competition*, pages 72–82, 2002.

[36] Takuya Hashimoto, Sachio Hiramatsu, and Hiroshi Kobayashi. Development of the face robot SAYA for rich facial expressions. In *2006 International Conference on Mechatronics and Automation*, pages 25–30. IEEE, 2006.

[37] Takuya Hashimoto, Sachio Hiramatsu, and Hiroshi Kobayashi. Dynamic display of facial expressions on the face robot made by using a life mask. In *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*, pages 521–526. IEEE, 2008.

[38] Zhanpeng He and Matei Ciocarlie. Morph: Design co-optimization with reinforcement learning via a differentiable hardware model proxy. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7764–7771. IEEE, 2024.

[39] Marcel Heisler and Christian Becker-Asano. An iPhone Pro is All You Need: Mimicking Facial Expressions on an Android Robot Head. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1347–1351. IEEE, 2025.

[40] Jonathan Hiller and Hod Lipson. Automatic design and manufacture of soft robots. *IEEE Transactions on Robotics*, 28(2):457–466, 2011.

[41] Yuhang Hu, Boyuan Chen, Jiong Lin, Yunzhe Wang, Yingke Wang, Cameron Mehlman, and Hod Lipson. Human-robot facial coexpression. *Science Robotics*, 9 (88):eadi4724, 2024.

[42] Yuhang Hu, Jiong Lin, Judah Allen Goldfeder, Philippe M Wyder, Yifeng Cao, Steven Tian, Yunzhe Wang, Jingran Wang, Mengmeng Wang, Jie Zeng, et al. Learning realistic lip motions for humanoid face robots. *Science Robotics*, 11(110):eadx3017, 2026.

[43] Simon Huber, Roi Poranne, and Stelian Coros. Designing actuation systems for animatronic figures via globally optimal discrete search. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021.

[44] Hisashi Ishihara, Yuichiro Yoshikawa, and Minoru Asada. Realistic child robot "affetto" for understanding the caregiver-child attachment relationship that guides the child development. In *2011 ieee international conference on development and learning (icdl)*, volume 2, pages 1–5. IEEE, 2011.

[45] Sharfin Islam, Zhanpeng He, and Matei Ciocarlie. Task-Based Design and Policy Co-Optimization for Tendon-driven Underactuated Kinematic Chains. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12016–12023. IEEE, 2024.

[46] Kazuko Itoh, Hiroyasu Miwa, Massimiliano Zecca, Hideaki Takanobu, Stefano Roccella, Maria Chiara Carrozza, Paolo Dario, and Atsuo Takanishi. *Mechanical design of emotion expression humanoid robot we-4rii*.

Springer, 2006.

[47] Xiaozhong Ji, Chuming Lin, Zhonggan Ding, Ying Tai, Junwei Zhu, Xiaobin Hu, Donghao Luo, Yanhao Ge, and Chengjie Wang. Realtalk: Real-time and realistic audio-driven face generation with 3d facial prior-guided identity alignment network. *arXiv preprint arXiv:2406.18284*, 2024.

[48] Kenji Kaneko, Fumio Kanehiro, Mitsuharu Morisawa, Kanako Miura, Shin'ichiro Nakaoka, and Shuuji Kajita. Cybernetic human HRP-4C. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pages 7–14. IEEE, 2009.

[49] Kento Kawaharazuka, Shunnosuke Yoshimura, Temma Suzuki, Kei Okada, and Masayuki Inaba. Design Optimization of Wire Arrangement With Variable Relay Points in Numerical Simulation for Tendon-Driven Robots. *IEEE Robotics and Automation Letters*, 9(2): 1388–1395, 2023.

[50] Kento Kawaharazuka, Kei Okada, and Masayuki Inaba. Robot Design Optimization with Rotational and Prismatic Joints using Black-Box Multi-Objective Optimization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4571–4577. IEEE, 2024.

[51] Hiroshi Kobayashi and Fumio Hara. Study on face robot for active human interface-mechanisms of face robot and expression of 6 basic facial expressions. In *Proceedings of 1993 2nd IEEE International Workshop on Robot and Human Communication*, pages 276–281. IEEE, 1993.

[52] Hiroshi Kobayashi, T Tsuji, and K Kikuchi. Study of a face robot platform as a kansei medium. In *2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies*, volume 1, pages 481–486. IEEE, 2000.

[53] Hiroshi Kobayashi, Yoshiro Ichikawa, Masaru Senda, and Taichi Shiiba. Realization of realistic and rich facial expressions by face robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 2, pages 1123–1128. IEEE, 2003.

[54] Milin Kodnongbua, Ian Good, Yu Lou, Jeffrey Lipton, and Adriana Schulz. Computational design of passive grippers. *ACM Transactions on Graphics (TOG)*, 41(4): 2–12, 2022.

[55] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[56] Dong-Wook Lee, Tae-Geun Lee, B So, Moosung Choi, Eun-Cheol Shin, K Yang, Moon-Hong Baek, Hong-Seok Kim, and Ho-Gil Lee. Development of an android for emotional expression and human interaction. In *Seventeenth world congress the international federation of automatic control, Seoul*, 2008.

[57] J. P. Lewis, Ken ichi Anjyo, Taehyun Rhee, Mengjie Zhang, Frédéric H. Pighin, and Zhigang Deng. Practice and Theory of Blendshape Facial Models. In *Eurographics*, 2014.

[58] Boren Li, Hang Li, and Hangxin Liu. Driving Animatronic Robot Facial Expression From Speech. *arXiv preprint arXiv:2403.12670*, 2024.

[59] Jian Li, Sheldon Andrews, Krisztian G Birkas, and Paul G Kry. Task-based design of cable-driven articulated mechanisms. In *Proceedings of the 1st Annual ACM Symposium on Computational Fabrication*, pages 1–12, 2017.

[60] Peizhen Li, Longbing Cao, Xiao-Ming Wu, Runze Yang, and Xiaohan Yu. X2C: A Dataset Featuring Nuanced Facial Expressions for Realistic Humanoid Imitation. *arXiv preprint arXiv:2505.11146*, 2025.

[61] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1, 2017.

[62] Wenji Li, Zhaojun Wang, Ruitao Mai, Pengxiang Ren, Qinchang Zhang, Yutao Zhou, Ning Xu, JiaFan Zhuang, Bin Xin, Liang Gao, et al. Modular design automation of the morphologies, controllers, and vision systems for intelligent robots: a survey. *Visual Intelligence*, 1(1):2, 2023.

[63] Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. Understanding Embodied Reference with Touch-Line Transformer. In *ICLR*, 2023.

[64] Chyi-Yeu Lin, Chang-Kuo Tseng, Wei-Chung Teng, Wei-Chen Lee, Chung-Hsien Kuo, Hung-Yan Gu, Kuo-Liang Chung, and Chin-Shyurng Fahn. The realization of robot theater: Humanoid robots and theatric performance. In *2009 International Conference on Advanced Robotics*, pages 1–6. IEEE, 2009.

[65] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13847–13858, 2025.

[66] Hod Lipson and Jordan B Pollack. Automatic design and manufacture of robotic lifeforms. *Nature*, 406 (6799):974–978, 2000.

[67] Chang Liu, Qunfen Lin, Zijiao Zeng, and Ye Pan. Emoface: Audio-driven emotional 3d face animation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 387–397. IEEE, 2024.

[68] Xiaofeng Liu, Yizhou Chen, Jie Li, and Angelo Cangelosi. Real-time robotic mirrored behavior of facial expressions and head motions based on lightweight networks. *IEEE Internet of Things Journal*, 10(2):1401–1413, 2022.

[69] 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*, 40:3850–3865, 2024.

[70] Xin Lu, Chuanqing Zhuang, Chenxi Jin, Zhengda Lu, Yiqun Wang, Wu Liu, and Jun Xiao. LSF-Animation: Label-Free Speech-Driven Facial Animation via Implicit Feature Representation. *arXiv preprint arXiv:2510.21864*, 2025.

[71] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines. *ArXiv*, abs/1906.08172, 2019.

[72] Ingo Lütkebohle, Frank Hegel, Simon Schulz, Matthias Hackel, Britta Wrede, Sven Wachsmuth, and Gerhard Sagerer. The bielefeld anthropomorphic robot head "Flobi". In *2010 IEEE international conference on robotics and automation*, pages 3384–3391. IEEE, 2010.

[73] Pragna Mannam, Xingyu Liu, Ding Zhao, Jean Oh, and Nancy Pollard. Design and control co-optimization for automated design iteration of dexterous anthropomorphic soft robotic hands. In *2024 IEEE 7th International Conference on Soft Robotics (RoboSoft)*, pages 332–339. IEEE, 2024.

[74] Vittorio Megaro, Bernhard Thomaszewski, Damien Gauge, Eitan Grinspun, Stelian Coros, and Markus H Gross. ChaCra: An Interactive Design System for Rapid Character Crafting. In *Symposium on Computer Animation*, pages 123–130, 2014.

[75] Shin'ichiro Nakaoka, Fumio Kanehiro, Kanako Miura, Mitsuharu Morisawa, Kiyoshi Fujiwara, Kenji Kaneko, Shuuji Kajita, and Hirohisa Hirukawa. Creating facial motions of Cybernetic Human HRP-4C. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pages 561–567. IEEE, 2009.

[76] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024.

[77] Celia Nieto Agraz, Pascal Hinrichs, Marco Eichelberg, and Andreas Hein. Is the robot spying on me? a study on perceived privacy in telepresence scenarios in a care setting with mobile and humanoid robots. *International Journal of Social Robotics*, 17(3):363–377, 2025.

[78] Jun-Ho Oh, David Hanson, Won-Sup Kim, Young Han, Jung-Yup Kim, and Ill-Woo Park. Design of android type humanoid robot Albert HUBO. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1428–1433. IEEE, 2006.

[79] Jia Pan, Sachin Chitta, and Dinesh Manocha. FCL: A general purpose library for collision and proximity queries. In *2012 IEEE international conference on robotics and automation*, pages 3859–3866. IEEE, 2012.

[80] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023.

[81] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023.

[82] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.

[83] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025.

[84] Ziqiao Peng, Wentao Hu, Junyuan Ma, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Hui Tian, Jun He, Hongyan Liu, and Zhaoxin Fan. SyncTalk++: High-Fidelity and Efficient Synchronized Talking Heads Synthesis Using Gaussian Splatting. *arXiv preprint arXiv:2506.14742*, 2025.

[85] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025.

[86] Francesco Penta, Cesare Rossi, and Sergio Savino. Analysis of suitable geometrical parameters for designing a tendon-driven under-actuated mechanical finger. *Frontiers of Mechanical Engineering*, 11(2):184–194, 2016.

[87] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[88] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1173–1182, 2021.

[89] Charles Schaff, David Yunis, Ayan Chakrabarti, and Matthew R Walter. Jointly learning to construct and control agents using deep reinforcement learning. In *2019 international conference on robotics and automation (ICRA)*, pages 9798–9805. IEEE, 2019.

[90] Qincheng Sheng, Wei Tang, Hao Qin, Yujie Kong, Haokai Dai, Yiding Zhong, Yonghao Wang, Jun Zou, and Huayong Yang. A review on an AI-driven face robot for human-robot expression interaction. *Science China Technological Sciences*, 68(10):2010301, 2025.

[91] Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. SAM Audio: Segment Anything in Audio. *arXiv preprint arXiv:2512.18099*, 2025.

[92] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024.

[93] Tao Sun, Bo Wang, and Xinming Huo. Knowledge-driven automated design of industrial robots: A unified graph-based framework with multi-engine reasoning. *Advanced Engineering Informatics*, 69:103995, 2026.

[94] Bernhard Thomaszewski, Stelian Coros, Damien Gauge, Vittorio Megaro, Eitan Grinspun, and Markus Gross. Computational design of linkage-based characters. *ACM Transactions on Graphics (TOG)*, 33(4):1–9, 2014.

[95] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024.

[96] Tingwu Wang, Yuhao Zhou, Sanja Fidler, and Jimmy Ba. Neural graph evolution: Towards efficient automatic robot design. *arXiv preprint arXiv:1906.05370*, 2019.

[97] Wu Weiguo, Men Qingmei, and Wang Yu. Development of the humanoid head portrait robot system with flexible face and expression. In *2004 IEEE International Conference on Robotics and Biomimetics*, pages 757–762. IEEE, 2004.

[98] Bowen Wu, Chaoran Liu, Carlos T Ishi, Takashi Minato, and Hiroshi Ishiguro. Retargeting human facial expression to human-like robotic face through neural network surrogate-based optimization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4724–4730. IEEE, 2024.

[99] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[100] Jie Xu, Tao Du, Michael Foshey, Beichen Li, Bo Zhu, Adriana Schulz, and Wojciech Matusik. Learning to fly: computational controller design for hybrid uavs with reinforcement learning. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[101] Jie Xu, Tao Chen, Lara Zlokapa, Michael Foshey, Wojciech Matusik, Shinjiro Sueda, and Pulkit Agrawal. An End-to-End Differentiable Framework for Contact-Aware Robot Design. In *Robotics: Science & Systems*, 2021.

[102] Zhuoxiong Xu, Xuanchen Li, Yuhao Cheng, Fei Xu, Yichao Yan, and Xiaokang Yang. SingingBot: An Avatar-Driven System for Robotic Face Singing Performance. *arXiv preprint arXiv:2601.02125*, 2026.

[103] Dongsheng Yang, Wataru Sato, Qianying Liu, Takashi Minato, Shushi Namba, and Shin'ya Nishida. Optimizing facial expressions of an android robot effectively: a bayesian optimization approach. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 542–549. IEEE, 2022.

[104] Dongsheng Yang, Qianying Liu, Wataru Sato, Takashi Minato, Chaoran Liu, and Shin'ya Nishida. HAPI: A Model for Learning Robot Facial Expressions from Human Preferences. *arXiv preprint arXiv:2503.17046*, 2025.

[105] Sha Yi, Xueqian Bai, Adabhav Singh, Jianglong Ye, Michael T Tolley, and Xiaolong Wang. Co-Design of Soft Gripper with Neural Physics. *arXiv preprint arXiv:2505.20404*, 2025.

[106] Dong Zhang, Jingwei Peng, Yuyang Jiao, Jiayuan Gu, Jingyi Yu, and Jiahao Chen. ExFace: Expressive Facial Control for Humanoid Robots with Diffusion Transformers and Bootstrap Training. *arXiv preprint arXiv:2504.14477*, 2025.

[107] Ran Zhang, Thomas Auzinger, Duygu Ceylan, Wilmot Li, and Bernd Bickel. Functionality-aware retargeting of mechanisms to 3D shapes. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[108] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023.

[109] 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.

[110] Yuang Zhang, Junqi Cheng, Haoyu Zhao, Jiaxi Gu, Fangyuan Zou, Zenghui Lu, and Peng Shu. Shoulder-Shot: Generating Over-the-Shoulder Dialogue Videos. *arXiv preprint arXiv:2508.07597*, 2025.

[111] Zongzheng Zhang, Jiawen Yang, Ziqiao Peng, Meng Yang, Jianzhu Ma, Lin Cheng, Huazhe Xu, Hang Zhao, and Hao Zhao. Morpheus: A Neural-driven Animatronic Face with Hybrid Actuation and Diverse Emotion Control. *arXiv preprint arXiv:2507.16645*, 2025.

[112] Allan Zhao, Jie Xu, Mina Konaković-Luković,

Josephine Hughes, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. Robogrammar: graph grammar for terrain-optimized robot design. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[113] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020.

[114] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. Interactive conversational head generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[115] Lifeng Zhu, Weiwei Xu, John Snyder, Yang Liu, Guoping Wang, and Baining Guo. Motion-guided mechanical toy modeling. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012.

[116] Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. INFP: Audio-driven interactive head generation in dyadic conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10667–10677, 2025.

[117] Yongtong Zhu, Lei Li, Iggy Qian, WenBin Zhou, Ye Yuan, Qingdu Li, Na Liu, and Jianwei Zhang. Awakening Facial Emotional Expressions in Human-Robot. *arXiv preprint arXiv:2510.23059*, 2025.

[118] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards Metrical Reconstruction of Human Faces. *European Conference on Computer Vision*, 2022.