DUPLEXAVATAR: UNIFIED 3D MOTION UNDERSTANDING AND GENERATION FOR FULL-DUPLEX SPOKEN DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

An interactive avatar must understand behavioral context and generate appropriate 3D motion during real-time spoken interaction. This requires behavioral understanding, multimodal response generation, and motion realization to operate within a shared conversational state. We introduce DUPLEXAVATAR, a unified full-duplex speech–language–motion model that maintains speech, language, behavior plans, and holistic 3D motion within a shared causal state, supporting both behavior understanding and multimodal response generation. At the core of DUPLEXAVATAR is the Executable Behavior Plan, a structured and executable representation of behavioral intent that specifies what to do, where and when it applies, and what must remain unchanged. The plan connects semantic decisions to controllable motion realization and provides an explicit interface for localized intervention. Structured operations are executed by a typed executor, while open-ended requests are realized through masked residual flow. Before commitment, geometry-based checks and motion-only readback can trigger a bounded local correction; only uncommitted future motion remains revisable, while committed motion is re-encoded into the conversational state. We further compile SPEAR, a multi-source, multi-view supervision suite that provides motion events, executable plans, geometry-grounded queries, and controlled counterfactual edits, together with an intervention-based benchmark for controllable behavior realization. Experiments under matched evaluation settings show consistent improvements across speech-conditioned motion generation, motion understanding, and behavioral control, achieving 73% human preference, 91.8% Event F1, and 78.95% constraint success. Together, DUPLEXAVATAR and SPEAR provide an inspectable path from behavioral intent to realized 3D motion within full-duplex spoken interaction.

1 INTRODUCTION

Natural spoken interaction is inherently embodied: speech, facial activity, posture, and hand gestures unfold together and continuously adapt to the conversation. An interactive avatar therefore requires two reciprocal capabilities: understanding visible behavior in context and generating 3D behavior that realizes its spoken response. Full-duplex speech models such as BayLing-Duplex enable listening while speaking and interruption within a shared streaming process (Fang et al., 2026), while co-speech and joint speech–motion models generate increasingly expressive holistic behavior (Yi et al., 2023; Liu et al., 2024; Zhang et al., 2025; Ma et al., 2026; Nagano et al., 2026; Lim et al., 2026). However, these two directions remain only partially connected: behavioral intent is typically latent, making it difficult for the dialogue model to explicitly interpret, control, revise, and verify the motion it produces.

To address this gap, we introduce **DUPLEXAVATAR**, a unified full-duplex speech–language–motion model that maintains speech, language, behavior plans, and holistic 3D motion within a shared causal conversational state. Regional motion streams for the face, upper body, lower body, and hands are integrated with the shared backbone, allowing the same model to support both behavior understanding and multimodal response generation. Motion can therefore be read as part of the conversational context rather than treated only as a final generated output.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

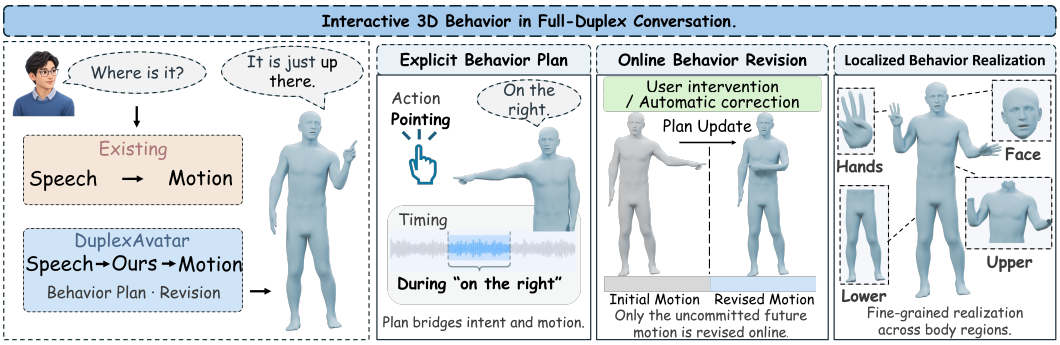

Figure 1: **Full-duplex avatar interaction with executable behavior control.** Left: DUPLEXAVATAR jointly listens, plans, speaks, and moves through a shared causal backbone, an Executable Behavior Plan, and a full-duplex commit boundary. Right: the plan explicitly specifies action and timing, supports revision of only uncommitted future motion under user intervention, and enables localized realization across body regions; pre-commit verification can trigger at most one local correction.

The key interface between these two directions is an **Executable Behavior Plan (EBP)**. An EBP explicitly represents what behavior to execute, where and when it applies, how it should be realized, and what must remain unchanged. Specifically, the plan connects semantic decisions to controllable motion realization: structured operations are handled by a typed executor, while open-ended behavioral requests are realized through masked residual flow over a speech-driven motion base. Before commitment, motion-only readback and geometric checks verify motion-observable constraints and can trigger at most one localized correction. Thus, the same representation supports behavior understanding, generation, localized intervention, and verification.

For online interaction, DUPLEXAVATAR separates committed motion from the uncommitted future. Plan-based revisions affect only the buffered future motion, while committed motion remains fixed. When an interruption occurs, pending speech, plan, and motion can be discarded together and generation resumes from the updated conversational context. After commitment, the executed motion is re-encoded into the shared state, so subsequent behavior is conditioned on what the avatar has actually performed.

To support this interface, we compile SPEAR (**S**peech-anchored **P**lans, **E**vents, **A**ttributes, and **R**ealization), a multi-source, multi-view supervision suite derived from verified speech–motion, motion–language, and motion-edit resources together with geometry-grounded supervision. It provides motion events, executable plans, SMPL-X-grounded queries, and controlled counterfactual edits, together with an intervention-based benchmark for controllable behavior realization. Experiments under matched evaluation settings demonstrate improvements across all three evaluated capabilities: DUPLEXAVATAR achieves 73% human preference for speech-conditioned motion generation, 91.8% Event F1 for motion understanding, and 78.95% constraint success for behavioral control.

In summary, our main contributions are that we

- introduce **DUPLEXAVATAR**, a unified full-duplex speech–language–motion model that maintains speech, language, and holistic 3D motion within a shared causal conversational state, supporting both behavior understanding and multimodal response generation.
- propose **Executable Behavior Plans**, a structured and executable representation of behavioral intent that connects semantic decisions to controllable motion realization, while supporting localized intervention, motion-based verification, bounded correction, and executed-state feedback.
- compile and derive **SPEAR**, a multi-source, multi-view supervision suite that provides motion events, executable plans, geometry-grounded queries, and controlled counterfactual edits, together with an intervention-based benchmark for evaluating controllable behavior realization.

2 Related Work

Full-duplex spoken models. Speech-language models increasingly operate directly on audio and generate spoken responses (Zeng et al., 2024). BayLing-Duplex organizes native full-duplex listening and speaking on a common causal timeline (Fang et al., 2026), while DuplexSLA jointly decodes speech, language, and a structured action channel on a shared chunk timeline (Zhang et al., 2026a). These models demonstrate unified streaming generation across multiple output channels and provide a foundation for richer forms of interactive multimodal behavior.

Joint speech and motion generation. Holistic co-speech systems model facial, body, and hand motion from speech (Yi et al., 2023; Liu et al., 2024; Zhang et al., 2025). ViBES introduces a speech–language–behavior model with modality experts, joint language–movement planning, and streaming body-action directives (Zhang et al., 2026b), while A^2-LLM connects conversational audio with avatar generation (Hu et al., 2026). Motion-Omni jointly produces speech and full-body motion, whereas DyaPlex and FacePlex study full-duplex speech with body or facial motion (Ma et al., 2026; Nagano et al., 2026; Lim et al., 2026). StreamTalk closes the streaming motion loop with a generate–retrieve–refine cycle and key-pose anchoring for long-horizon drift (Zhang et al., 2026c). Together, these works extend spoken interaction beyond audio-only responses toward synchronized facial and whole-body behavior.

Motion-language modeling and controllable editing. Reciprocal motion–language modeling supports text-to-motion, motion captioning, and instruction following (Guo et al., 2022a;b; Jiang et al., 2023; Wang et al., 2026; Wu et al., 2025b; Zhu et al., 2026). The Language of Motion (LoM) factorizes holistic movement into regional token streams (Chen et al., 2025; 2026). MotionFix introduces natural-language motion refinement pairs (Athanasiou et al., 2024), while SimMotionEdit, MotionReFit, OmniME, and MotionMERGE improve semantic alignment, locality, or multi-granular control (Li et al., 2025; Jiang et al., 2025; Shi et al., 2026; Wu et al., 2026). Building on these complementary directions, DuplexAvatar brings reciprocal motion–language modeling and controllable behavior realization into full-duplex spoken interaction. Its Executable Behavior Plan provides an explicit, editable representation of behavioral intent, while the commit mechanism restricts online revision to uncommitted future motion.

3 Method

3.1 Problem Formulation

At interaction block b, the model observes incoming user speech $\mathbf{a}^{\text{in}}_{\leq b}$ and the committed conversation history $\mathcal{H}_{<b}$. It produces assistant text $\mathbf{y}_b$, outgoing speech $\mathbf{a}^{\text{out}}_b$, an Executable Behavior Plan $\mathcal{P}_b$, and holistic motion

$$\mathbf{M}_b = \left(\mathbf{M}^{\text{face}}_b, \mathbf{M}^{\text{upper}}_b, \mathbf{M}^{\text{lower}}_b, \mathbf{M}^{\text{hands}}_b\right). \tag{1}$$

We model the initial response as

$$p_\theta\left(\mathbf{y}_b, \mathbf{a}^{\text{out}}_b, \mathcal{P}_b, \mathbf{M}_b \mid \mathbf{a}^{\text{in}}_{\leq b}, \mathcal{H}_{<b}\right), \tag{2}$$

where θ denotes the model parameters.

A short future block remains pending until playback commits it. Let $\mathbf{M}^{\text{base}}_b$ and $\mathcal{P}^{\text{pending}}_b$ denote the stored base motion and plan for this uncommitted block. Given an optional user intervention ι_b, the pending plan is updated as $\mathcal{P}^{\text{upd}}_b = U(\mathcal{P}^{\text{pending}}_b, \iota_b)$, where U denotes the plan update operation. The corresponding local revision is modeled as

$$p_\theta\left(\mathbf{M}^{\text{rev}}_b \mid \mathbf{M}^{\text{base}}_b, \mathcal{P}^{\text{upd}}_b, \mathcal{H}_{<b}\right). \tag{3}$$

Revision is restricted to the uncommitted block, while committed motion becomes part of the interaction history and conditions subsequent responses.

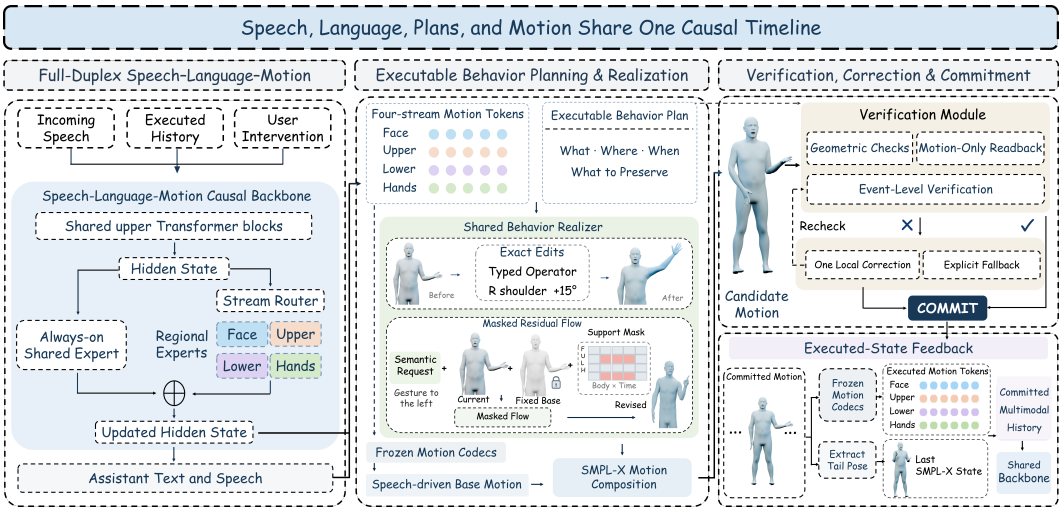

Figure 2: **DUPLEXAVATAR architecture.** Incoming speech, assistant language, outgoing voice, plans, and four motion streams share a causal block clock and a backbone with an always-on shared expert and four regional motion experts. The speech-driven motor path produces a base motion, while the plan-gated behavior path enters a shared realizer through typed exact operators or masked residual flow. Motion-derived geometry and motion-only readback support at most one local correction. Unresolved candidates fall back to the base motion with a failure status. Only committed motion feeds back through codec re-encoding and continuous motion state, while uncommitted future motion remains revisable.

3.2 UNIFIED CAUSAL SPEECH AND MOTION MODELING

We build on a pretrained full-duplex autoregressive speech model and retain its native $0.8\,\mathrm{s}$ interaction clock. Plan and motion are predicted through additional depth passes that reuse the upper transformer parameters within the same interaction block. Speech, language, plans, and motion therefore share a causal state and commitment boundary.

We augment each of the upper eight transformer layers with an always-active shared expert E_s^ℓ and four regional motion experts $\{E_r^\ell\}_{r\in\mathcal{R}}$, where $\mathcal{R} = \{\mathrm{face}, \mathrm{upper}, \mathrm{lower}, \mathrm{hands}\}$. The motion-stream identity deterministically selects the corresponding regional expert, while speech, text, and plan positions use only the shared path. The layer output is

$$\widetilde{\mathbf{h}}_i^\ell = \mathbf{h}_i^\ell + E_s^\ell(\mathbf{h}_i^\ell) + \mathbf{1}_{i\in\mathcal{M}}\,\alpha_{r_i}^\ell E_{r_i}^\ell(\mathbf{h}_i^\ell), \tag{4}$$

where r_i denotes the regional stream of motion position i and $\alpha_{r_i}^\ell$ is a learned residual scale. Each regional expert uses learned input and output projections to bridge the BayLing and LoM hidden spaces, and all four are initialized from the same pretrained LoM FFN before specializing on their respective motion streams. Shared attention and the shared expert preserve cross-region and cross-modal interaction.

Holistic motion is represented by four pretrained and frozen LoM codecs for the face, upper body, lower body, and hands (Chen et al., 2025; 2026). The four streams retain their native face and body rates while remaining aligned to the shared interaction clock. All motion streams condition on committed motion history. The face stream uses outgoing-audio embeddings, whereas the body motion pass uses causally available assistant-audio tokens together with the continuous executed tail, providing prosodic and stylistic cues. Cross-region behavior is coordinated by the plan-conditioned realizer.

Region-specific embeddings and output projections map codec indices to the shared hidden space. Motion is decoded into SMPL-X space for continuous composition and geometric supervision, then re-encoded after realization so subsequent blocks condition on the committed motion.

3.3 Executable Behavior Planning and Realization

An Executable Behavior Plan contains up to eight independently timed events,

$$\mathcal{P}_b = (e_{b,1}, \ldots, e_{b,N_b}), \qquad 0 \le N_b \le 8. \tag{5}$$

Each event specifies

$$\begin{aligned} e_{b,i} = \big(&\texttt{goal}_i, \texttt{region}_i, \texttt{target}_i, \texttt{interval}_i, \\ &\texttt{operation}_i, \texttt{axis}_i, \texttt{magnitude}_i\big). \end{aligned} \tag{6}$$

The plan also specifies execution mode and preservation flags. `target` identifies a joint, regional joint set, expression coefficient, or root field. Event intervals are quantized at $40\,\mathrm{ms}$ within each $0.8\,\mathrm{s}$ block, and a constrained grammar limits field values and serialization length. An empty plan leaves the base motion unchanged.

The behavior realizer composes a speech-driven base motion with a plan-conditioned residual:

$$\widehat{\mathbf{M}}_b = \mathcal{C}\big(\mathbf{M}_b^{\mathrm{base}}, \Delta\mathbf{M}_b(\mathcal{P}_b), \mathbf{S}_b\big)\,, \tag{7}$$

where $\mathbf{S}_b$ is the union of the events' field and temporal supports, including transition ramps. Composition multiplies selected local rotations and adds selected expression or root-velocity increments. Fields outside the support are copied from the base motion:

$$\Big(\widehat{\mathbf{M}}_b\Big)_{\neg\mathbf{S}_b} = \big(\mathbf{M}_b^{\mathrm{base}}\big)_{\neg\mathbf{S}_b}\,. \tag{8}$$

This guarantee holds in the composed parameter space. Descendant joints may still move in world space after an ancestor rotates, and codec round trips need not be lossless.

For exact requests, a typed executor applies local joint rotations, expression-coefficient increments, and root-velocity increments with temporal ramps. Their magnitudes are represented in degrees, percentage points, and centimetres per second, respectively. For open semantic goals, a masked residual-flow head predicts a modification conditioned on the plan, the current motion, and a fixed reference. A shared execution pass reads the full plan and motion sequences and produces frame-aligned states $\mathbf{h}_b^{\mathrm{exec}}$ without direct access to the raw request. The original base candidate remains fixed as the reference throughout correction, while the current motion changes after a failed realization. For a relative rotation,

$$R_{\mathrm{goal}} = R_{\mathrm{ref}}\Delta R_{\mathcal{P}}, \qquad \Delta R_{\mathrm{repair}} = R_{\mathrm{current}}^{\top} R_{\mathrm{goal}}. \tag{9}$$

We train the residual-flow head with endpoint prediction. For selected fields, let $Z_t = (1-t)\epsilon + t\Delta\mathbf{M}^{\star}$. The realization objective is

$$\mathcal{L}_{\mathrm{real}} = \mathbb{E}_{t,\epsilon}\Big[\big\|\mathbf{S}_b \odot \big(f_\theta(Z_t, t, \mathbf{h}_b^{\mathrm{exec}}) - \Delta\mathbf{M}^{\star}\big)\big\|_{\mathrm{grp}}^2\Big]\,, \tag{10}$$

where the grouped norm averages valid elements within each active field group and then across groups. Boundary-velocity and available geometric supervision regularize the composed result, while identity residuals and deterministic copying preserve fields outside the support.

Motion-only readback and bounded correction. A realized candidate is re-encoded into the four motion streams and paired with geometric evidence from its continuous SMPL-X motion. Without access to the original request, speech, or intended plan, a fresh local depth pass reads back motion-observable plan fields and descriptive semantics. Geometric checks verify directly measurable goals and exact typed operations, while open semantic goals are checked only through observable region, temporal support, and realized kinematics; non-observable communicative intent is left unverified rather than treated as a mismatch. Only verified mismatches trigger at most one corrective regeneration, after which unresolved violations fall back to the base motion.

Pending revision and executed-state feedback. A user edit updates the pending plan and invokes the same realizer from the stored base motion. Assistant text, outgoing speech, plan events, and composed motion remain pending until commitment. A barge-in cancels pending outputs and in-flight results while retaining consumed user audio. Commitment updates the motion state used by the next block, so subsequent behavior conditions on the motion actually retained by the runtime.

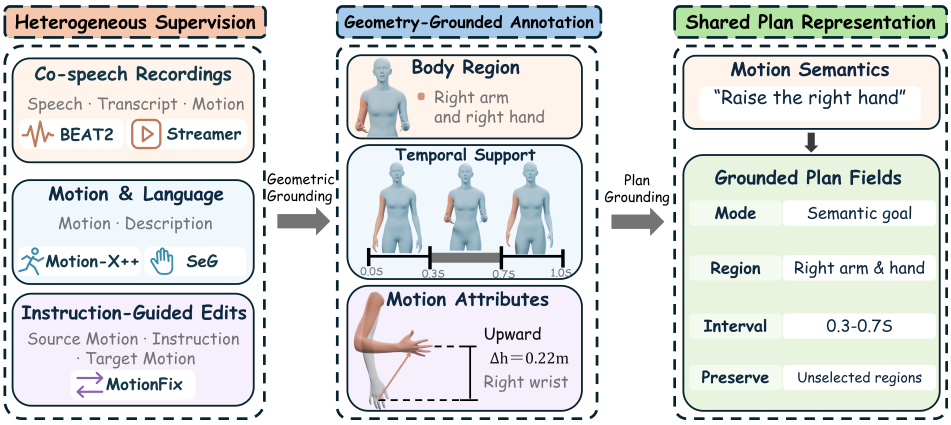

Figure 3: **SPEAR supervision construction.** Heterogeneous speech–motion, motion–language, and motion-edit sources are converted into geometry-grounded event annotations and a shared plan representation. The resulting supervision supports reciprocal motion understanding, generation, readback, and controlled intervention.

3.4 LEARNING OBJECTIVE

The motion-adaptation objective combines plan prediction, motion-only readback, four-stream motion prediction, residual realization:

$$\mathcal{L}_{\text{adapt}} = \lambda_{\text{plan}}\mathcal{L}_{\text{plan}} + \lambda_{\text{read}}\mathcal{L}_{\text{readback}} + \lambda_{\text{mot}}\mathcal{L}_{\text{motion}} + \lambda_{\text{real}}\big(\mathcal{L}_{\text{real}} + 0.1\mathcal{L}_{\text{boundary}} + \mathcal{L}_{\text{geom}}\big). \tag{11}$$

$\mathcal{L}_{\text{plan}}$ supervises structured plan fields and goal tokens, including direct supervision of the existing start and end tick logits. Train-only same-prefix pairs provide paired supervision for interval, target, axis, and magnitude without introducing a separate prediction head. $\mathcal{L}_{\text{readback}}$ supervises motion-observable plan fields and descriptive motion semantics from motion-only inputs.

Natural co-speech records supervise the base motion, while controlled edits and aligned motion pairs supervise residual realization. Request-to-plan and speech-to-plan tasks supervise intended behavior, while motion-to-plan supervision is restricted to fields and descriptive semantics that are observable from motion.

4 SPEAR SUPERVISION

SPEAR provides multi-source supervision for fine-grained embodied behavior by organizing verified speech–motion, motion–language, and motion-edit resources under a shared plan representation. Starting from 57,044 source records, we derive 87,573 motion windows, 140,298 rule-grounded behavior plans, and 3.86M geometry-grounded queries. Among the source records, 22,589 contain reliably aligned speech. The same source records are converted into reciprocal task views for speech- and plan-conditioned motion generation, motion description, motion-to-plan recovery, and plan-conditioned realization. Speech-specific tasks use only records with reliable speech and temporal alignment.

We derive geometry-grounded supervision from continuous SMPL-X motion to associate semantic behavior with observable body regions, joints, temporal support, and motion attributes. These annotations provide structured evidence for executable-plan grounding and motion-only readback, connecting semantic intent to the motion that realizes it.

Controlled operators further construct counterfactual edits by modifying selected plan fields while preserving the remaining source motion. Natural motion-refinement pairs provide complementary supervision for open-ended edits (Athanasiou et al., 2024). All task views derived from the same speaker and source recording are kept in the same split to prevent cross-view leakage.

Table 1: **Quantitative comparison.** Results under matched input, output, and body coverage.

Motion generation Method	Motion Smoothness ↑	Speech–Motion Alignment ↑	Human Pref. (%) ↑
EMAGE	1.84	0.53	9
SemTalk	1.91	0.74	12
StreamTalk	3.14	0.64	6
DUPLEXAVATAR	**3.42**	**0.79**	**73**
Motion understanding Method	Event F1 (%) ↑	Body-Part Accuracy (%) ↑	Attribute Accuracy (%) ↑
MotionGPT	83.3	83.3	33.3
MG-MotionLLM	50.0	41.7	8.3
DUPLEXAVATAR	**91.8**	**94.3**	**86.7**
Behavioral control Method	Constraint Success (%) ↑	Preservation Drift (mm) ↓	Target Error (mm) ↓
SimMotionEdit	59.13	55.19	151.14
MotionReFit	51.68	56.72	197.85
DUPLEXAVATAR	**78.95**	**8.64**	**18.73**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on speaker-disjoint splits, with all test speakers and source recordings excluded from training and validation. Task views derived from the same source recording remain in the same split to avoid cross-view leakage. Speech-specific evaluations use only records with reliable speech and temporal alignment.

Because existing methods support different subsets of the complete setting, we use task-specific matched protocols rather than a single common baseline pool. For speech-conditioned motion generation, all methods receive the same assistant speech or audio conditioning and are compared under matched body coverage. For motion understanding, all methods receive the same motion inputs and are evaluated under a common event-level protocol covering event recognition, body-part grounding, and behavioral attributes. For behavioral control, methods receive the same source motion and edit instruction, and are evaluated on both target satisfaction and preservation outside the specified spatiotemporal support.

Human evaluation complements the automatic motion-generation metrics by asking raters to select the generated motion that best matches the semantic content of the speech. Full-duplex interaction is evaluated separately from these matched offline comparisons using streaming latency, real-time execution, and barge-in recovery.

5.2 COMPARISON EXPERIMENTS

Quantitative comparison. We compare DUPLEXAVATAR with representative baselines on holistic motion generation, motion understanding, and precise behavioral control. For motion generation, we compare with EMAGE, SemTalk, and StreamTalk (Liu et al., 2024; Zhang et al., 2025; 2026c) under matched body coverage and conditioning. We evaluate motion smoothness and speech–motion alignment. Motion smoothness is computed as the inverse of the average third-order temporal variation of upper-body joint motion, while speech–motion alignment is measured using Beat Consistency.

As shown in Table 1, DUPLEXAVATAR achieves the strongest performance across all three motion-generation metrics. In particular, it improves speech–motion alignment to 0.79 while maintaining a motion smoothness of 3.42. The human study shows a larger perceptual difference, with 73% of preferences assigned to DUPLEXAVATAR. These results indicate that jointly modeling the conversational state and holistic motion improves both temporal coordination with speech and the semantic appropriateness of the generated behavior.

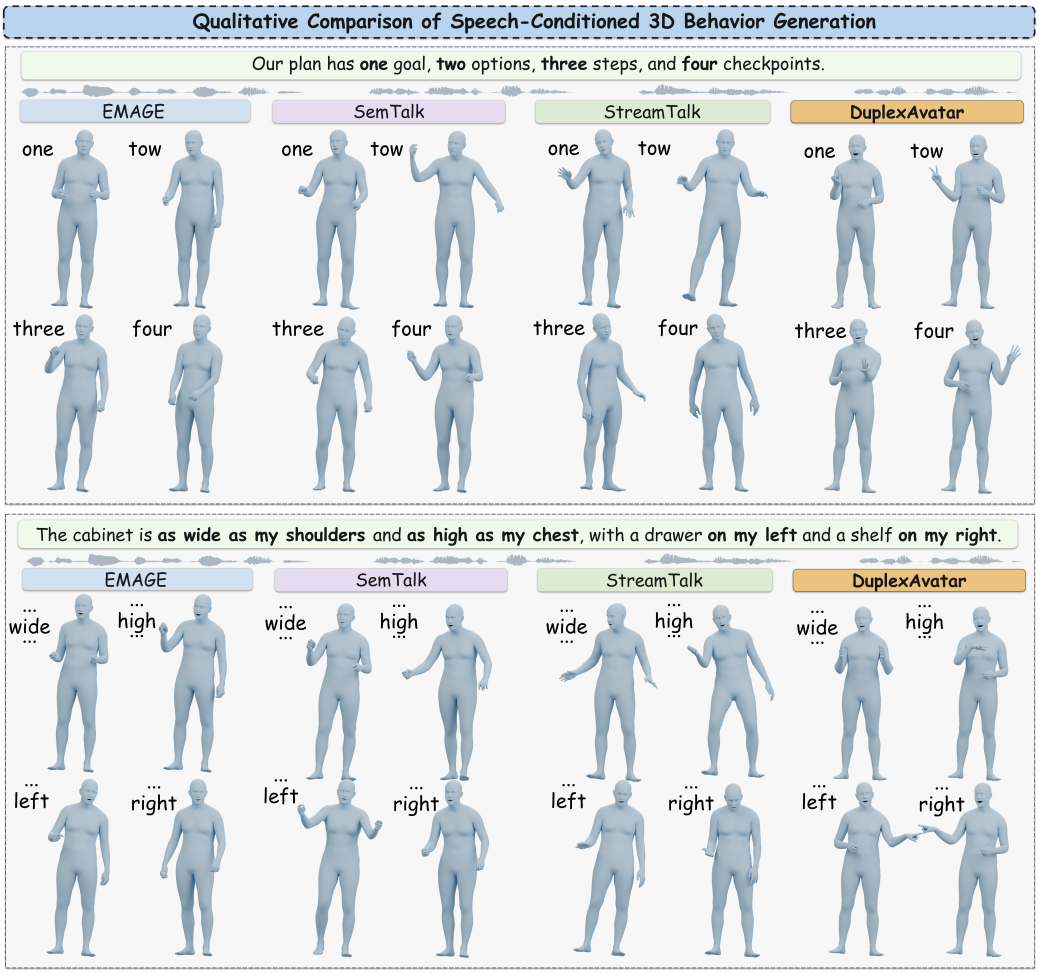

Figure 4: **Qualitative comparison of speech-conditioned 3D behavior.** All systems receive the same speech waveform. EMAGE, SemTalk, StreamTalk, and DUPLEXAVATAR generate corresponding SMPL-X motion. Representative frames illustrate speech–motion timing and semantic correspondence.

For motion understanding, we compare with MotionGPT and MG-MotionLLM (Jiang et al., 2023; Wu et al., 2025a). Given motion alone, we evaluate structured behavior understanding using Event F1, Body-Part Accuracy, and Attribute Accuracy, measuring whether the model correctly recognizes the behavioral event, grounds it to the responsible body region, and recovers the associated motion attributes.

DUPLEXAVATAR reaches 91.8% Event F1 and 94.3% Body-Part Accuracy, showing that the shared motion representation preserves both event-level semantics and regional grounding. The larger gain in Attribute Accuracy, where DUPLEXAVATAR achieves 86.7%, further indicates that the model recovers fine-grained motion-observable properties rather than only coarse action categories.

For precise behavioral control, we compare with representative motion-editing methods under matched source motion and instructions. We evaluate whether the requested target is achieved and whether motion outside the specified spatiotemporal support is preserved. This setting directly measures the locality and executability of the behavior representation.

DUPLEXAVATAR achieves a constraint success rate of 78.95%, compared with 59.13% for SimMotionEdit and 51.68% for MotionReFit. Preservation drift decreases to 8.64 mm and target error to 18.73 mm. The simultaneous improvement in target accuracy and preservation indicates that the model localizes the requested change while retaining motion outside the specified support.

Table 2: **Ablation study.** Effects of planning, realization, and verification components.

Variant	Motion Smoothness ↑	Constraint Success (%) ↑	Preservation Drift (mm) ↓	Target Error (mm) ↓
w/o EBP	3.30	31.58	30.42	76.85
w/o Regional Experts	2.84	65.79	14.73	32.16
w/o Typed Executor	3.35	52.63	18.26	47.91
w/o Masked Residual Flow	2.96	60.53	15.84	36.72
w/o Readback Correction	3.37	68.42	11.36	27.45
Full model	**3.42**	**78.95**	**8.64**	**18.73**

Qualitative comparison. We compare DUPLEXAVATAR with EMAGE, SemTalk, and StreamTalk (Liu et al., 2024; Zhang et al., 2025; 2026c). All systems receive the same speech waveform. Methods requiring textual input additionally receive the same transcript. StreamTalk uses a fixed speaker identity and retrieval database. Figure 4 visualizes representative moments to compare speech–motion timing, body coverage, and semantic correspondence.

Full-duplex interaction. We evaluate the online interaction capability of DUPLEXAVATAR using time to first audio (TTFA), time to first motion (TTFM), real-time factor (RTF), and barge-in recovery. TTFA and TTFM measure the responsiveness of streaming speech and motion outputs, while RTF measures whether the joint pipeline operates in real time. Barge-in recovery evaluates whether the model continues to accept user speech during its own response and can cancel pending speech and motion before continuing from the updated conversational context.

5.3 ABLATION STUDY

We ablate the components most directly associated with behavior planning, motion specialization, realization, and verification. *w/o EBP* removes the Executable Behavior Plan and generates motion directly from the conversational context. *w/o Regional Experts* removes the four region-specific motion experts and retains only the shared expert in the augmented layers. *w/o Typed Executor* removes deterministic execution of structured operations and routes all behavior realization through the learned residual path. *w/o Masked Residual Flow* removes the learned residual realization path for open-ended behaviors while retaining the EBP and Typed Executor. *w/o Readback Correction* disables the bounded corrective pass triggered by motion-only semantic readback and geometric verification. We evaluate all variants using motion smoothness, constraint success, preservation drift, and target error.

The ablations show complementary contributions from planning, specialization, realization, and verification. Removing the EBP causes the largest drop in constraint success, confirming the importance of an explicit behavioral interface. Removing the regional experts degrades both motion quality and control accuracy, while removing either realization branch reduces target accuracy. Disabling readback correction further increases preservation drift and target error, showing that the bounded verification loop corrects residual failures after the initial realization.

6 CONCLUSION

We presented DUPLEXAVATAR, a unified full-duplex speech–language–motion model that connects behavioral understanding, intent, and visible 3D motion within a shared causal conversational state. Executable Behavior Plans provide an explicit interface for motion understanding, controllable realization, intervention, and verification before commitment. Executed motion is fed back into the dialogue state, while SPEAR provides the semantic, geometric, and interventional supervision needed to learn this interface. On the three main evaluation settings, DUPLEXAVATAR reaches 73% human preference for speech-conditioned motion generation, 91.8% Event F1 for motion understanding, and 78.95% constraint success for behavioral control. These results support the use of explicit behavioral plans for making 3D motion a controllable part of full-duplex spoken interaction.

REFERENCES

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3D human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11. ACM, 2024.

Changan Chen, Juze Zhang, Shrinidhi K. Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3D human motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6200–6211, 2025.

Changan Chen, Juze Zhang, Shrinidhi K. Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. Language of motion v2. GitHub repository, 2026. URL `https://github.com/Juzezhang/language_of_motion`.

Qingkai Fang, Shoutao Guo, and Yang Feng. Bayling-duplex: Native full-duplex speech dialogue with a single autoregressive LLM. *arXiv preprint arXiv:2606.14528*, 2026.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *European Conference on Computer Vision*, pp. 580–597. Springer, 2022b.

Xiaolin Hu, Hang Yuan, Xinzhu Sang, Binbin Yan, Zhou Yu, Cong Huang, and Kai Chen. A^2-LLM: An end-to-end conversational audio avatar large language model. *arXiv preprint arXiv:2602.04913*, 2026.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20067–20079, 2023.

Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22735–22745, 2025.

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. SimMotionEdit: Text-based human motion editing with motion similarity prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27827–27837, 2025.

Habin Lim, Jae-Ho Lee, Hah Min Lew, Ji-Su Kang, and Gyeong-Moon Park. FacePlex: Full-duplex joint speech-facial motion generation for conversational avatars. *arXiv preprint arXiv:2606.30145*, 2026.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1144–1154, 2024.

Chengqian Ma, Wei Tao, Haoyu Zhang, and Yiwen Guo. Motion-omni: End-to-end joint speech and full-body motion for spoken dialogue. *arXiv preprint arXiv:2609.04250*, 2026.

Koki Nagano, Hongyu Liu, Seonwook Park, Tianye Li, Amrita Mazumdar, Christian Jacobsen, Shengze Wang, Michael Stengel, Rajarshi Roy, Ka Chun Cheung, Simon See, and Shalini De Mello. DyaPlex: Full-duplex speech-motion model for dyadic interaction. *arXiv preprint arXiv:2606.03874*, 2026.

Zhenwu Shi, Jingyu Gong, Peiwei Wang, Xingzan Wang, Tianwen Qian, Wenxi Li, Yuan Fang, Jiao Xie, Lizhuang Ma, and Shaohui Lin. Omni-supervised motion editing: Balancing change and invariance through positive-negative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Dan Xu, and Shixiang Tang. MotionGPT-2: A general-purpose motion-language model for motion generation and understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(8):11027–11041, 2026.

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27849–27858, 2025a.

Bizhu Wu, Jinheng Xie, Wenting Chen, Zhe Kong, Jianfeng Ren, Linlin Shen, Ruibin Bai, and Rong Qu. MotionMERGE: A multi-granular framework for human motion editing, reasoning, generation, and explanation. *arXiv preprint arXiv:2605.18956*, 2026.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-Agent: A conversational framework for human motion generation with LLMs. In *International Conference on Learning Representations*, 2025b.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3D human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 469–480, 2023.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Haoyang Zhang, Jun Chen, Donghang Wu, Yuxin Li, Yuxin Zhang, Xiangyu Tony Zhang, Che Liu, Qingjian Lin, Yizhou Peng, Hexin Liu, Eng Siong Chng, Chao Yan, Boyong Wu, Yechang Huang, Xuerui Yang, and Fei Tian. DuplexSLA: A full-duplex spoken language model with synchronized speech, language, and action. *arXiv preprint arXiv:2605.20755*, 2026a.

Juze Zhang, Changan Chen, Xin Chen, Heng Yu, Tiange Xiang, Ali Sartaz Khan, Shrinidhi K. Lakshmikanth, and Ehsan Adeli. ViBES: A conversational agent with behaviorally-intelligent 3D virtual body. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39994–40008, 2026b.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Ziqiang Dang, Jianqiang Ren, Liefeng Bo, and Zhigang Tu. SemTalk: Holistic co-speech motion generation with frame-level semantic emphasis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13761–13771, 2025.

Xiangyue Zhang, Jianfang Li, Jiaxu Zhang, Kaixing Yang, and Steven Hoi. StreamTalk: Streaming co-speech gesture generation with key-pose anchoring. In *European Conference on Computer Vision*, 2026c.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *International Conference on Learning Representations*, 2026.