MindFlow: Harmonizing Cognitive Semantics and Acoustic Dynamics for Facial Animation Generation in Dyadic Conversations

Hejia Chen[1⋆], Haoxian Zhang[2], Xu He[2], Xiaoqiang Liu[2], Pengfei Wan[2], Shoulong Zhang[3]✉, and Shuai Li[1,3]✉

[1] Beihang University
[2] Kling Team, Kuaishou Technology
[3] Zhongguancun Laboratory
https://harryxd2018.github.io/MindFlow/

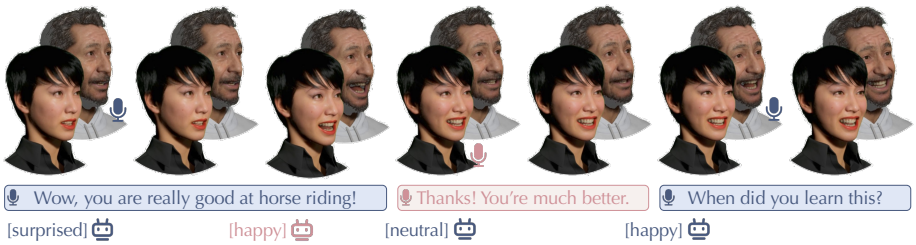

Fig. 1: Grounded in the Ventral-Dorsal dual-pathway cognitive model, **MindFlow** introduces a novel framework for streaming facial animation in dyadic conversations, under which digital avatars simultaneously *perceive* conversational emotions while reflexively *synchronizing* with acoustic rhythms, naturally yielding interactions that are both semantically rich and physically fluid.

Abstract. Generating lifelike facial animation for dyadic conversations requires reconciling high-level cognitive intent with precise low-level motor reflexes, yet existing methods fall short in the semantic understanding of dialogue context and in precise dynamic control. In this paper, we propose **MindFlow**, a dual-pathway generative framework inspired by the Ventral-Dorsal pathway model in neuroscience, which decouples generation into two collaborative streams, thereby harmonizing deep semantic reasoning with fine-grained control. In the Ventral module, we transform the conventional Sentence-Action approach into a novel Chunk-State approach that models raw acoustic streams as a context-aware, evolving emotional state chain, capturing subtle paralinguistic nuances and mid-utterance emotional shifts missed by sentence-level modeling. The Dorsal module features a conditional autoregressive flow matching

⋆ This work was conducted during the internship at Kling Team, Kuaishou Technology

network for high-fidelity facial motion, driven by high-frequency acoustic cues and modulated by emotion states, plus a Selective Acoustic Injector for adaptive audio gating to ensure robustness in talking-and-listening dynamics without interference. Extensive experiments demonstrate that MindFlow achieves superior semantic appropriateness and motion naturalness compared to state-of-the-art baselines.

Keywords: Facial Animation · Multimodal LLM · Flow Matching

1 Introduction

Creating realistic digital avatars capable of fluid, natural conversational interactions is a longstanding fundamental goal within the computer vision and graphics community [26, 30, 32]. In particular, generating natural facial dynamics for avatars acting as active conversational interlocutors has garnered significant research attention across multiple fields, including education [34], entertainment [21], and social companionship [12]. However, despite substantial recent advancements, facial animations of both talking and listening behaviors produced by existing approaches often appear stiff and semantically hollow, due to insufficient semantic understanding of dialogue and limited fine-grained controllability, which limits avatar realism and user engagement.

Compellingly, in neuroscience, a well-established Ventral-Dorsal dual-pathway model [16] explains that natural conversation relies on two distinct neural pathways working in parallel: 1) a fast, reflexive *Dorsal pathway* that processes prosodic information and maps acoustic signals directly to articulatory motor areas, and 2) a slow, deliberative *Ventral pathway* that consumes high-level cognitive resources and decodes complex semantic and emotional information. However, existing computational approaches have yet to fully reconcile this functional duality in conversational facial animation generation. Traditional audio-driven methods [29, 30, 32, 51] seek to directly map listening behaviors from audio signals, mimicking the Dorsal pathway's functionality, yet they neglect the explicit semantic understanding modeled by the Ventral pathway. This results in animations that are visually reactive but semantically ungrounded and hollow. While recent works have attempted to address this gap by leveraging large language models (LLMs) as proxies for the Ventral pathway to decode sentence-level semantics from the text modality [26, 45], their sentence-to-action mapping approch suffers from two critical unresolved limitations: 1) *prosodic information loss*: relying solely on text modality inevitably discards critical paralinguistic cues and emotional semantics inherent in the raw conversational audio; and 2) *coarse granularity*: sentence-level segmentation is inherently coarse, which fails to semantically guided fine-grained temporal dynamics of human expression during conversation.

To address the aforementioned limitations, we propose **MindFlow**, a novel dual-stream framework for high-fidelity conversational facial animation generation in dyadic interactions. Drawing direct inspiration from the Ventral-Dorsal dual-pathway model [16] in neuroscience, we decouple conversation-driven facial

animation generation into two complementary streams: a cognitive semantic-understanding stream (dubbed Mind) and a reflexive flow-matching-based generation stream (dubbed Flow), implemented via our proposed Ventral module and Dorsal module, respectively.

The proposed **Ventral module** extracts high-level semantic and emotional intent from the evolving dialogue context by incorporating multimodal LLMs (MLLMs). To overcome the problems of coarse granularity and prosodic information loss, rather than using the conventional Sentence-Action approach, we design a novel chunk-state approach that encodes the stream of fixed-window raw dialogue audio chunks and extracts corresponding emotion states. Additionally, we devise a streaming Chain-of-State mechanism that dynamically aggregates historical audio chunks and their corresponding emotion states as an expanding context for the MLLM, enabling temporally consistent reasoning across continuous audio streams. Our Ventral module can preserve delicate paralinguistic cues and encode acoustic context into temporally evolving, continuous emotion states, thereby replacing the static, rigid sentence summaries used in previous work.

Complementarily, the proposed **Dorsal module** aims to mimic the sensorimotor reflex system to generate high-fidelity, physically plausible facial motions. To support continuous, variable-length inference and mitigate boundary artifacts [30, 32], we utilize a conditional autoregressive flow matching architecture [43]. However, existing methods often implement a static fusion of dyadic audio signals, diluting effective acoustic features. This diminishes the capability of the audio signal to guide facial motion generation, consequently weakening the avatar's expressiveness during both listening and speaking phases. To explicitly address this critical gap, our key innovation in this module is a novel *Selective Acoustic Injector*. By adaptively filtering high-frequency raw audio features, this component selects the most relevant audio signals to drive facial motions based on the conversational state. Consequently, it guarantees precise, audio-synchronized lip movements during speech and focuses more on the conversational partner's voice during listening phases to generate semantically aligned reactive expressions.

Extensive experiments confirm that our novel dual-stream approach for facial animation generation in the dyadic conversation setting improves lip synchronization and expression accuracy compared with state-of-the-art methods. In summary, our main contributions are as follows:

1. We propose MindFlow, a novel dual-stream framework inspired by the Ventral-Dorsal pathway model, designed to address the problem of rigid, hollow facial expressions in dyadic conversation animation.
2. We establish a Chunk-State approach that enables MLLMs in the Ventral module to analyze the interlocutor's dynamic emotion states from audio streams, effectively preserving audio information and enabling continuous, fine-grained control for expression generation.
3. We design a Selective Acoustic Injector in the Dorsal module, which adaptively gates listening and talking dynamics to generate high-quality streaming facial motions seamlessly across both talking and listening states.

2 Related Work

Dorsal-Stream-Based Generation. Early studies [25,29,35] and [4,7,13,30,32,49, 50] build upon talking-head architectures [6,11,14,15,43,44,47,48] by introducing listening branches to extract reactive cues directly from audio signals. These approaches primarily rely on learning low-level correlations between acoustic patterns and facial dynamics. Recent advancements [5,20,36] further distill video generation models into autoregressive frameworks to enable online streaming synthesis. While these methods effectively model the temporal causality required for low-latency reflexes, they function similarly to a motor system without a cognitive center. By focusing solely on signal-level mapping, they often fail to capture the long-term semantic coherence and evolving emotion states inherent in dyadic conversations.

Dorsal-Ventral Dual-Stream-Based Generation. To incorporate high-level understanding, recent works leverage LLMs to guide facial animation. Some methods [22,29] fine-tune LLMs [2] to directly predict reactions, but often suffer from scenario-specific biases due to limited training data. A more prevalent strategy, adopted by [19,26,38,41,45], utilizes external LLMs to analyze dialogue content and generate predefined behaviors, such as specific head nods or gaze shifts. We term this the Sentence-Action approach. Although this approach ensures semantic fitness, it suffers from fundamental limitations inherent to its design: relying primarily on the textual modality discards crucial prosodic and emotional cues, while sentence-level reasoning results in coarse-grained, unsynchronized reactions. Similarly, recent holistic conversational frameworks [18,28,31,39] integrate LLMs to synthesize comprehensive multimodal responses. However, these systems predominantly prioritize the *active talking* phase—generating content and motion during the avatar's turn—often overlooking the continuous, semantically grounded non-verbal feedback required during *active listening*. Unlike these approaches, our framework shifts to a Chunk-State approach, ensuring continuous semantic guidance for both talking and listening roles.

3 Method

3.1 Problem Formulation

We formulate the task of dyadic facial animation generation from the perspective of an interlocutor a engaged in a symmetric interaction with another interlocutor b. Drawing inspiration from the natural operation of the human sensory system, which processes continuous stimuli sequentially, we formulate this task as a causal, streaming generation problem, which naturally diverges from existing non-causal approaches that rely on accessing the full conversational context [26, 30, 32, 51]. At any time step t, the system generates the facial motion M_a^t for interlocutor a, conditioning solely on the historical information available up to t. The generation process G is modeled as:

$$M_a^t = G(A_a^{\leq t}, A_b^{\leq t}, S_a^{<t}), \tag{1}$$

where $A_a^{\leq t}$ and $A_b^{\leq t}$ denote continuous raw audio streams, and $S_a^{<t}$ represents the evolving emotion state derived from the interaction history. By adhering to this streaming causality, our formulation mirrors the continuous information flow of biological communication, where future contexts remain inherently inaccessible. Under this biomimetic constraint, the primary challenge is to simultaneously ensure instantaneous audio-visual alignment for dynamic conversational behavior while maintaining the long-term coherence of emotion states.

3.2 Framework: Ventral-Dorsal Dual Pathway

Inspired by the biological Ventral-Dorsal dual-pathway model, which separately processes semantic comprehension and sensory-motor reflexes, we construct the MindFlow framework (Fig. 2). Specifically, MindFlow incorporates a Ventral module and a Dorsal module to emulate these respective functions, where the two modules collaborate seamlessly to ensure the avatar's facial dynamics are both contextually appropriate and consistently engaging, effectively overcoming the issue of hollow expressions common in traditional audio-driven methods.

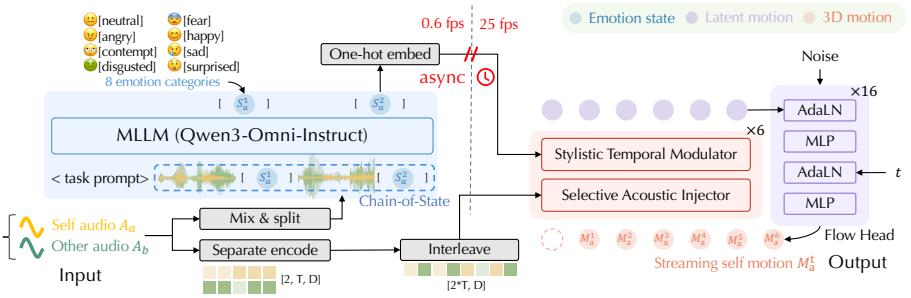

Fig. 2: Inspired by the Ventral-Dorsal dual-pathway model, **MindFlow** generates life-like conversational facial animations (both listening and speaking) driven by continuous dialogue audio A_a and A_b. 1) The Ventral module functions as a cognitive semantic perceiver, leveraging an MLLM with a streaming *Chain-of-State* to decode evolving emotion states chunk-by-chunk, conditioned on the historical cognitive trajectory. 2) The Dorsal module functions as a reflexive sensory-motor executor. It incorporates a *Selective Acoustic Injector* that uses motion queries to adaptively gate unmixed audio streams, and a *Stylistic Temporal Modulator* that injects the Ventral emotion states as semantic guidance. These combined features condition an autoregressive flow-matching backbone for the continuous generation of high-fidelity, synchronous facial motions.

To bridge these two modules, we introduce a novel *Chunk-State approach*. Previous methods [26, 45] typically adopt a *Sentence-Action approach*, which performs behavior reasoning based on sentence-level textual content. To plan

proper movements, they often force the reasoning model to output verbose motion descriptions burdened with complex spatial and temporal details, leading to temporally mismatched and coarse-grained responses. In contrast, our approach operates directly on raw audio streams at a much finer chunk-level granularity. Because it reasons at the scale of short chunks, our Ventral module naturally achieves more precise alignment without needing to explicitly predict temporal details. Furthermore, aligning with the biological dual-pathway model, we argue that intricate spatial expression details should not be pre-planned cognitively. Instead, the Ventral module acts solely as the cognitive foundation, continuously ingesting audio chunks to decode and update a fluid emotion state. Concurrently, the Dorsal module functions as the sensory-motor executor. Conditioned on the Ventral module's semantic guidance and instantaneous acoustic cues, it takes over the generation of fine spatial details, translating them into high-fidelity, synchronous reflexes including accurate lip articulation, rhythmic head movements, and dynamic expressions. Formally, this dual-pathway process at time step t can be formulated as:

$$\begin{cases} S_a^k = \text{Ventral}(A_a^{\leq k}, A_b^{\leq k}, S_a^{<k}) \\ M_a^t = \text{Dorsal}(A_a^{\leq t}, A_b^{\leq t}, M_a^{<t}, S_a^{\lfloor t/w \rfloor}) \end{cases}, \tag{2}$$

where $k = \lfloor t/w \rfloor$ denotes the discrete chunk index, and w represents the chunk window size. Through this chunk-state collaboration, MindFlow bypasses the stiffness of sentence-level textual planning, guaranteeing that the avatar not only comprehends the semantic flow to react reasonably but also remains physically "alive" and highly responsive at every moment.

3.3 Ventral Module: Streaming Cognitive State Modeling

The Ventral module functions as the cognitive anchor of the MindFlow framework, emulating the biological Ventral pathway responsible for semantic and emotional comprehension. Its primary objective is to maintain an emotion state derived from the evolving conversation history. By operating on a coarse temporal scale, it distills the continuous audio streams into context-aware cognitive states via semantic reasoning.

Unlike the Sentence-Action approach [26, 45] that relies on sentence-level reasoning, our Chunk-State approach analyzes the interlocutor's emotional variations using fine-grained audio chunks as the fundamental unit. Operating directly on the audio modality offers two critical advantages. First, it avoids the inevitable loss of prosodic information that occurs when converting speech to textual transcripts. Second, the smaller temporal window of audio chunks significantly enhances the temporal granularity of the semantic guidance provided by the Ventral module. Consequently, we turn to MLLMs [8, 40] to perform direct semantic and emotional analysis natively within the audio modality.

However, existing MLLMs are predominantly tailored for turn-based question-answering tasks, posing a significant challenge for streaming continuous conversations. If we were to independently query the MLLM to analyze the emotion of

each individual chunk, the rich conversational context would be completely destroyed. Conversely, if we simply concatenated the current chunk with historical chunks as the input query, the MLLM would lack explicit awareness of its past emotion states, frequently leading to unstable and temporally inconsistent predictions for the current chunk. To resolve this dilemma, we propose a streaming Chain-of-State mechanism (Fig. 3) that explicitly models the evolving emotion state while simultaneously preserving the complete history of the interaction.

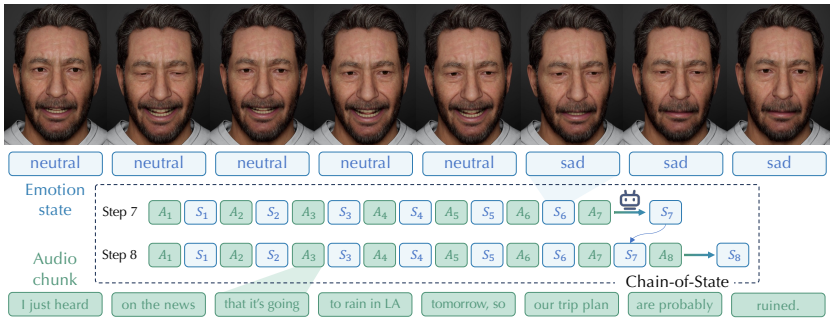

Fig. 3: Visualization of Ventral module reasoning and corresponding generated result. The MetaHuman plugin in Unreal Engine is utilized for rendering.

Formally, at step k, the Ventral module predicts the current state S_a^k conditioned not only on the current audio chunk A_k (as mixed chunk $A_a^{wk:wk+w}$ and $A_b^{wk:wk+w}$ from both interlocutors) but also on the history of past audio inputs and previously inferred states ($A_a^{\leq wk}, A_b^{\leq wk}, S_a^{<k}$). By progressively appending the predicted emotion state to the context alongside the audio stream, this formulation establishes an expanding memory of both the sensory input and the cognitive trajectory. These derived emotion states are concurrently transmitted to the Dorsal module to guide facial motion generation. Ultimately, this Chain-of-State mechanism effectively mitigates the inherent instability of memory-less MLLM inferences; by explicitly conditioning on the trajectory of past states, it preserves contextual continuity, ensuring the avatar's emotional expressions evolve smoothly, robustly, and coherently throughout the continuous interaction.

3.4 Dorsal Module: Reflexive Motion Generation

Acting as the sensory-motor executor of our framework, the Dorsal module emulates the biological Dorsal pathway, tasked with generating immediate subconscious reflexes driven by high-frequency acoustic cues. To achieve high-fidelity motion that is both responsive and physically coherent, we design a unified architecture comprising an autoregressive Transformer backbone tailored with mechanism-specific injectors, and a flow matching head for generating expressive and diverse facial motion [3, 43].

Stylistic Temporal Modulator. To ensure that the generated reflexes adhere to the emotion state obtained by the Ventral module, we introduce the Stylistic Temporal Modulator. Implemented as a specialized Transformer encoder layer, this module enforces semantic consistency while preserving temporal continuity. The emotion state S_a^k from the Ventral module is first embedded and appended to the input hidden motion sequence at step $t = wk$. We employ a masked causal attention mechanism where the emotion state embedding is visible to subsequent hidden motion $H_m^{wk:wk+w}$ as guidance, while preventing information leakage from future frames. Guided by this semantic-aware emotion state, the Dorsal module generates rich high-frequency dynamics, ensuring that the avatar's expressions during the conversation are both vivid and engaging.

Selective Acoustic Injector. Following the sensory-motor integration function of the biological Dorsal pathway, the Dorsal module takes the audio streams of both interlocutors (A_a and A_b) as input. Audio-driven methods [30,32,51] commonly adopt concatenation-based fusion to early-mix these distinct signals. Specifically, they concatenate the audio features and map them through an MLP prior to cross-attention injection, which can be formulated as:

$$F_{\text{concat}} = \text{Attn}(H_m, A_{\text{fuse}}, A_{\text{fuse}}), \quad \text{where } A_{\text{fuse}} = \text{MLP}(\text{Concat}(A_a, A_b)). \quad (3)$$

This early-mix strategy prematurely entangles disparate acoustic sources into a shared latent space. Since the attention mechanism is forced to query from this mixed context, it obscures the distinct functional boundaries required for talking articulation versus listening reaction, inevitably causing signal dilution.

To address this, we introduce the Selective Acoustic Injector, implemented as a modified Transformer decoder layer. Instead of early mixing, the audio features from both interlocutors are temporally interleaved to form a composite, source-independent acoustic context $A_{\text{ctx}} = \text{Interleave}(A_a, A_b)$. We formulate the injection process as a query-response mechanism where the motion history H_m actively attends to relevant acoustic cues:

$$F_{\text{inject}} = \text{Attn}(H_m, A_{\text{ctx}}, A_{\text{ctx}}) = \text{Softmax}\left(\frac{Q_m K_{\text{ctx}}^\top}{\sqrt{d}}\right) V_{\text{ctx}}. \quad (4)$$

By exposing the unmixed audio streams directly to the attention mechanism, this design allows the injector to learn a dynamic gating policy. Crucially, we do not impose any explicit supervision or speaker labels to direct this attention mechanism. Instead, the dynamic gating strategy acts as an emergent behavior optimized purely by the flow matching objective. As visualized in the attention heatmap (Fig. 4), the model exhibits a clear, unsupervised phase transition: while generating the motion of interlocutor a, the query predominantly locks onto the *self* track (A_a) for precise lip-sync during active speech, but spontaneously shifts focus to the *other* track (A_b) to trigger reactive behaviors during listening. This confirms that our injector naturally internalizes the turn-taking dynamics of dyadic conversation, dynamically filtering the acoustic context based solely on motion causality.

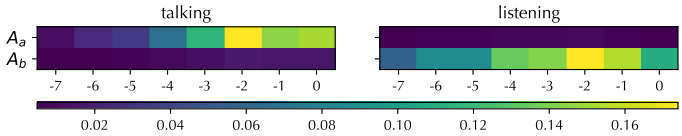

Fig. 4: Attention map showing motion query focusing on relevant audio.

Autoregressive Transformer Backbone. To effectively integrate the stylistic and acoustic representations, our backbone is constructed by alternately stacking the aforementioned modules. Specifically, the network comprises $L = 6$ identical blocks, where each block sequentially applies the Stylistic Temporal Modulator and the Selective Acoustic Injector. During training, the backbone operates under a teacher-forcing strategy, utilizing the historical ground-truth motion sequence $M_a^{<t}$ to predict the motion condition C_a^t for the current frame. Formally, we define the end-to-end conditioning generation process as:

$$C_a^t = \text{Backbone}\left(M_{<t}, S_a^t, A_{ctx}\right), \tag{5}$$

where the function Backbone$(\cdot)$ encapsulates the feature embedding, L alternating blocks, and a final linear projection. The output C_a^t provides the precise expression context required for the subsequent flow matching process.

Flow Matching Head. To capture the inherent expressive diversity of realistic interactions, we discard deterministic regression in favor of a generative flow matching approach [24]. We formulate facial motion generation as learning a transport map from a standard Gaussian distribution $\pi_0 = \mathcal{N}(0, I)$ to the data distribution π_1. The flow matching head, parameterized as an MLP v_θ, learns to predict the velocity field that drives the ordinary differential equation:

$$dZ_\tau = v_\theta(Z_\tau, \tau|C)d\tau, \tag{6}$$

where Z_τ is the motion state at timestep $\tau \in [0, 1]$, and C represents the conditioning context from the transformer backbone. During training, we optimize the flow matching objective:

$$\mathcal{L}_{flow} = \mathbb{E}_{\tau\sim\mathcal{U}[0,1], Z_0\sim\pi_0, Z_1\sim\pi_1}[\|v_\theta(Z_\tau, \tau|C) - (Z_1 - Z_0)\|^2]. \tag{7}$$

Unlike diffusion models [17] that require iterative denoising, flow matching learns straight-line trajectories, allowing us to approximate the solution using a simple 5-step Euler solver during inference. This efficiency is critical, enabling our Dorsal module to synthesize high-fidelity, non-deterministic facial details with a rapid responsiveness characteristic of the biological Dorsal pathway.

4 Experiments

4.1 Implementation Details

The Dorsal module is trained on a combination of public datasets totaling approximately 20 hours: HDTF [46] provides fundamental motion priors; VICOX[4] strengthens listening behavior and interaction modeling; and MEAD [37] and VICO [50] capture emotional talking and listening behaviors, respectively. We adopt a two-stage strategy: the model is first pre-trained on HDTF and VICOX for 90k steps to establish basic talking and conversational behaviors, then fine-tuned on MEAD and VICO for 30k steps to improve emotion awareness and expressiveness ($\sim$ 4 days in total). Both stages use the Adam optimizer with a batch size of 64, a peak learning rate of $1e^{-5}$ with a cosine schedule and a 1% warmup, and no weight decay. The audio encoder [1] is frozen throughout both stages. Facial movements are encoded as 51-dimensional ARKit blendshape coefficients for expressions and 3D Euler angles for head pose, extracted via MediaPipe [27] and FSA-Net [42], respectively.

During inference, the two modules run asynchronously. The Ventral module processes each 1.5 s audio chunk in 1.38 ± 0.10s, while the Dorsal module generates facial reactions at 25 FPS in real time, reusing S_a^k until the next state S_a^{k+1} arrives. The full system requires 59 GB VRAM and sustains real-time output over 2-minute sequences without memory growth.

4.2 Comparison with Prior Methods

Quantitative Evaluation

Metrics. We introduce a comprehensive set of evaluation metrics to systematically assess the quality of generated facial movements in both talking and listening states. To evaluate motion realism, we employ the Fréchet Distance (FD). For assessing lip synchronization during talking phases, we utilize SyncNet scores (SyncC and SyncD), while Mean Squared Error (MSE) is adopted to measure expression accuracy within the listening context. Notably, we adapt SyncNet [9, 23] from video-based synchronization evaluation to the 3D modality.

Talking Performance. We compare the talking performance of our method with previous approaches on the HDTF testset [46]. Specifically, we compare with talking head models EmoTalk and UniTalker [10, 33], which use the same motion representation as ours. Meanwhile, we re-train the dyadic methods [30,32] to align with our motion representation. As shown in Tab. 1a, MindFlow achieves state-of-the-art performance across all metrics, encompassing lip synchronization, facial expression fidelity, and head movement realism. Notably, our Dorsal module operates via autoregressive rollout to generate motion in a frame-by-frame streaming manner, supporting continuous inference for long conversational audio and overcoming the fixed-length limitations ($\sim 10s$) of previous methods.

[4] `https://project.mhzhou.com/vico`

Table 1: Quantitative comparison with SOTA methods. We evaluate performance on two distinct interaction modes: (a) talking generation and (b) listening responsiveness. Green and Cyan highlights denote the best and second-best results.

(a) Evaluation on talking state.

Method	SyncNet		FD ↓	
	SyncD ↓	SyncC ↑	Exp	Pose
EmoTalk	0.429	0.412	23.52	-
UniTalker	0.480	0.300	29.88	-
DualTalk	0.467	0.346	26.06	0.18
A2P	0.341	0.519	17.64	0.03
Ours	0.333	0.520	15.76	0.01

(b) Evaluation on listening state.

Method	FD↓		MSE↓	
	Exp	Pose	Exp	Pose
L2L	33.93	0.06	0.93	0.01
RLHG	39.02	0.07	0.86	0.01
DIM	23.88	0.06	0.70	0.01
DualTalk	22.27	0.05	0.58	0.01
A2P	14.24	0.03	0.34	0.01
Ours	13.86	0.03	0.30	0.01

Listening Performance. We follow the evaluation protocol of DualTalk to assess the quality of generated listening behaviors on the VICO testset. As shown in Tab. 1b, our model achieves lower scores in terms of FD and MSE, indicating that it generates more accurate and appropriate facial movements during the listening state. We attribute this improvement to: 1) the introduction of semantic guidance as a prior for motion generation, making the generated results more appropriate; and 2) the use of a generative network architecture, which is capable of producing more diverse facial motion patterns and effectively mitigates the over-smoothing issue commonly observed in discriminative models.

Qualitative Evaluation We further conduct a qualitative comparison to assess the proposed method from the perspectives of motion naturalness and lip-sync consistency. First, we adopt DualTalk and A2P as baseline Dorsal-only methods for comparison. Fig. 5 presents representative results generated by different methods over multiple dialogue turns, including alternating talking and listening states. Compared to DualTalk, whose head pose exhibits noticeable discrepancies in motion amplitude and frequency between talking and listening states, our method produces smoother and more natural head movements across different interaction modes. In contrast to A2P, which struggles to maintain stable facial expressions over extended durations, our approach is able to preserve controllable and coherent emotional expressions over long sequences. We attribute this limitation of A2P to its fixed-length diffusion transformer structure, which hinders the long-term consistency of facial motions via sliding-window inference, whereas the autoregressive flow matching generation structure in the Dorsal module enables more stable and reliable inference for long-duration facial animation.

Second, we compare with CustomListener[5], which leverages LLM for semantic understanding on sentence-level textual dialogue. As illustrated in Fig. 6, our

[5] Obtained from the CustomListener homepage.

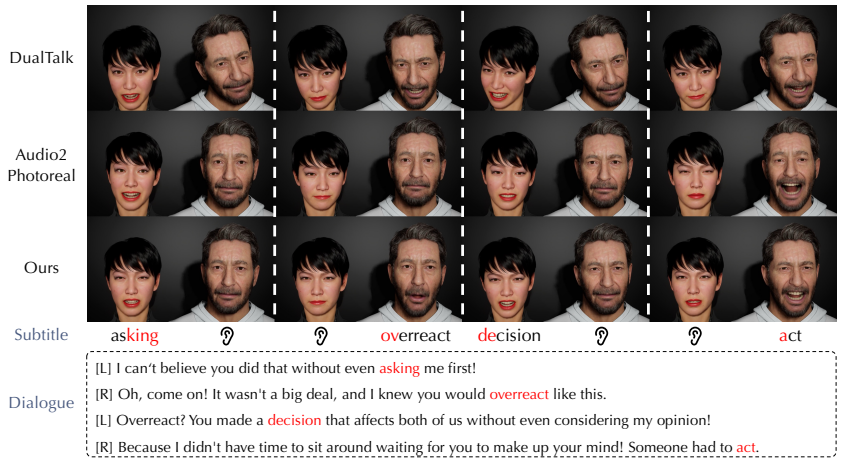

Fig. 5: Performance under conversational tension: MindFlow captures the intense atmosphere to exhibit semantically and temporally appropriate expressions, whereas baseline methods suffer from unstable and hollow expressions due to the boundary artifacts and a lack of semantic guidance.

approach generates appropriate, well-timed reactions (*i.e.*, a smile at the end). In contrast, CustomListener fails to produce noticeable reactions and exhibits static facial expressions, a limitation stemming from its coarse sentence-level action constraints. This indicates that our Chunk-State approach possesses a distinct advantage in temporal granularity over the Sentence-Action approach.

Finally, we conduct a perception study for a more comprehensive evaluation. We designed a survey utilizing a comparative format, where evaluators were asked to select the best generation results based on two criteria: the naturalness and fitness of the generated movements. To ensure a blind evaluation and prevent bias, the correspondence between the methods and the presented options was randomized. As illustrated in Fig. 7, out of the 24 collected evaluations, our model consistently outperforms previous methods in both naturalness and fitness.

4.3 Ablation Study

Semantic Guidance. We analyze the impact of the Ventral module's semantic guidance provided to the Dorsal module through quantitative ablation experiments. Specifically, on the VICO testset, we provide the Dorsal module with three distinct types of guidance signals: a random state, a fixed state (sentence-level GT), and the emotion state derived from the Ventral module. As shown in Tab. 2a, compared to random or fixed guidance, the evolving emotion state extracted from the dialogue by the Ventral module effectively guides the Dorsal module to generate more contextually appropriate facial expressions. This indi-

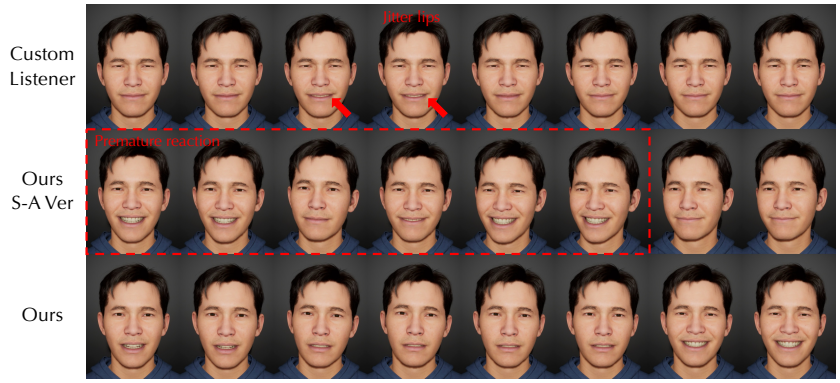

Fig. 6: Compared to the Sentence-Action approach, our proposed Chunk-State approach yields reactions with more appropriate timing and style.

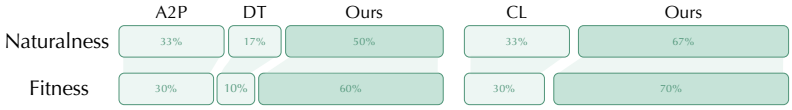

Fig. 7: Results of the perception study.

cates the guidance from the Ventral module is temporally dynamic, fine-grained, and aligned with the ground-truth facial expressions.

Chunk-State Approach. We compare our method against a Sentence-Action variant of our own model, which degrades the Chain-of-State mechanism to sentence-level reasoning. As shown in Fig. 6, although the variant can also yield appropriate emotional states, the resulting reactions are temporally unsynchronized and premature. This distinctly underscores the critical role of fine-grained semantic reasoning, which is further validated in Tab. 2a (*v.s.* S-A).

Table 2: Results of component-wise ablation study.

(a) Ventral module.

Method	FD↓		MSE ↓	
	Exp	Pose	Exp	Pose
Random	15.21	0.03	0.36	0.01
Fixed	14.15	0.03	0.32	0.01
S-A	14.39	0.03	0.33	0.01
Ours	13.86	0.03	0.30	0.01

(b) Selective Acoustic Injector (SAI).

Method	SAI	SyncD↓	SyncC↑
A2P	w/o	0.341	0.519
	w/	0.331	0.526
Ours	w/o	0.350	0.424
	w/	0.333	0.520

Reason Unit Length. The window size of the chunk w serves as a critical hyperparameter governing the trade-off between the granularity of emotion state and stability. To identify the optimal setting, we conduct a user study to assess emotion prediction accuracy (Acc) and perceptual synchronicity (pSync) under different chunk sizes. As shown in Fig. 8, while increasing the chunk size improves accuracy, which is likely due to the incorporation of richer temporal context, it degrades the perceptual synchronicity. To balance these trade-offs, we empirically set the chunk size to 1.5 seconds.

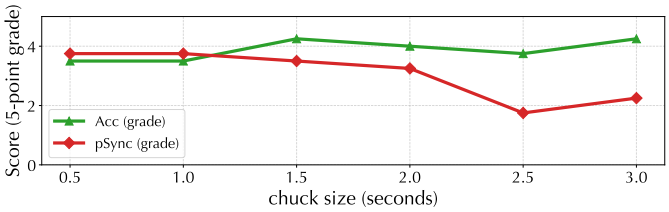

Fig. 8: Impact of chunk size.

Selective Acoustic Injector. We validate this injector by comparing our method and a retrained A2P [30] against two cross-ablated variants: (1) A2P augmented with our proposed injector, and (2) our method using the A2P acoustic module. As shown in Tab. 2b, incorporating the injector consistently improves lip-sync quality across architectures. This demonstrates its effectiveness in selectively modeling cross-temporal acoustic features and providing more informative guidance. Moreover, the superior performance of the A2P-augmented variant over our causal method aligns with the known advantage of bidirectional approaches, further validating the consistent effectiveness of the proposed injector.

5 Conclusion

We presented MindFlow, a novel dual-stream framework that bridges the gap between high-level cognitive reasoning and precise motor reflexes for streaming dyadic facial animation. By introducing a hierarchical Chunk-State approach, MindFlow overcomes the limitations of rigid sentence-level constraints. Ventral module leverages a streaming Chain-of-State mechanism to extract continuous, fine-grained emotional semantics directly from audio chunks. Concurrently, the Dorsal module employs a Selective Acoustic Injector and a flow matching head to generate high-fidelity, synchronous facial dynamics seamlessly across both talking and listening phases. Extensive experiments demonstrate that MindFlow significantly outperforms state-of-the-art methods in semantic appropriateness and motion naturalness, enabling more lifelike and engaging conversational avatars.

6 Limitations and Future Work

While MindFlow significantly advances audio-driven dyadic facial animation, it presents certain limitations that pave the way for future research. Primarily, our current framework relies exclusively on auditory inputs to infer the interlocutor's emotional and semantic states. In natural face-to-face interactions, visual information, *e.g.*, the interlocutor's eye contact, facial expressions, and body language, is equally crucial for conveying non-verbal intent and emotional nuances. Relying solely on the acoustic modality may occasionally limit the avatar's capacity to react to silent yet semantically rich behaviors. Therefore, a key direction for future work is to extend our framework to accommodate multimodal sensory inputs. By seamlessly integrating the interlocutor's visual cues alongside the raw audio streams, the framework can achieve a more holistic contextual understanding, enabling the generation of even more empathetic, accurate, and contextually appropriate facial responses.

Acknowledgments

This work was supported by the National Key R&D Program of China (2023YF F1203803), Zhongguancun Laboratory, and the National Natural Science Foundation of China (62502469, 62525204).

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: NeurIPS. pp. 12449–12460 (2020)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS. vol. 33, pp. 1877–1901 (2020)
3. Cen, Z., Pi, H., Peng, S., Shuai, Q., Shen, Y., Bao, H., Zhou, X., Hu, R.: Ready-to-React: Online reaction policy for two-character interaction generation. In: ICLR (2025)
4. Chatziagapi, A., Morency, L.P., Gong, H., Zollhöfer, M., Samaras, D., Richard, A.: Av-flow: Transforming text to audio-visual human-like interactions. In: ICCV. pp. 14270–14282 (2025)
5. Chen, B., Liu, H.: Dystream: Streaming dyadic talking heads generation via flow matching-based autoregressive model. arXiv preprint arXiv:2512.24408 (2025)
6. Chen, H., Zhang, H., Zhang, S., Liu, X., Zhuang, S., Wan, P., ZHANG, D., Li, S., et al.: Cafe-talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control. In: ICLR (2025)
7. Chen, M., Cui, L., Zhang, W., Zhang, H., Zhou, Y., Li, X., Tang, S., Liu, J., Liao, B., Chen, H., et al.: Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation. arXiv preprint arXiv:2508.19320 (2025)
8. Chen, Q., Cheng, L., Deng, C., Li, X., Liu, J., Tan, C.H., Wang, W., Xu, J., Ye, J., Zhang, Q., Zhang, Q., Zhou, J.: Fun-audio-chat technical report. arXiv preprint arXiv:2512.20156 (2025)

9. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV 2016 Workshops. pp. 251–263. Springer (2017)
10. Fan, X., Li, J., Lin, Z., Xiao, W., Yang, L.: Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In: ECCV. pp. 204–221 (2024)
11. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM ToG **40**(4) (2021)
12. Gao, Y., Dai, Y., Zhang, G., Guo, H., Hao, A., Li, S.: Effects of interaction modalities and emotional states on user's perceived empathy with an llm-based embodied conversational agent. International Journal of Human-Computer Studies **204**, 103585 (2025)
13. Guo, Y., Liu, X., Zhen, C., Yan, P., Wei, X.: ARIG: Autoregressive interactive head generation for real-time conversations. In: ICCV. pp. 12956–12965 (2025)
14. He, X., Huang, Q., Zhang, Z., Lin, Z., Wu, Z., Yang, S., Li, M., Chen, Z., Xu, S., Wu, X.: Co-speech gesture video generation via motion-decoupled diffusion model. In: CVPR. pp. 2263–2273 (2024)
15. He, X., Zhang, H., Chen, H., Zheng, C., Chen, L., Tang, S., Huang, J., Liu, X., Wan, P., Wu, Z.: From inpainting to editing: A self-bootstrapping framework for context-rich visual dubbing. arXiv preprint arXiv:2512.25066 (2025)
16. Hickok, G., Poeppel, D.: The cortical organization of speech processing. Nature reviews neuroscience **8**(5), 393–402 (2007)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020)
18. Hu, Y., Liu, R., Ren, Y., Yin, X., Li, H.: Unitalker: Conversational speech-visual synthesis. In: ACM MM. pp. 10248–10257 (2025)
19. Ji, X., Lin, C., Ding, Z., Tai, Y., Zhu, J., Hu, X., Luo, D., Ge, Y., Wang, C.: Realtalk: Real-time and realistic audio-driven face generation with 3d facial prior-guided identity alignment network. arXiv preprint arXiv:2406.18284 (2024)
20. Ki, T., Jang, S., Jo, J., Yoon, J., Hwang, S.J.: Avatar forcing: Real-time interactive head avatar generation for natural conversation. arXiv preprint arXiv:2601.00664 (2026)
21. Kyrlitsias, C., Michael-Grigoriou, D.: Social interaction with agents and avatars in immersive virtual environments: A survey. Frontiers in Virtual Reality **2**, 786665 (2022)
22. Lai, P., Zhong, W., Qin, Y., Ren, X., Wang, B., Li, G.: LLM-driven multimodal and multi-identity listening head generation. In: CVPR. pp. 10656–10666 (2025)
23. Li, C., Zhang, C., Xu, W., Lin, J., Xie, J., Feng, W., Peng, B., Chen, C., Xing, W.: LatentSync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. arXiv preprint arXiv:2412.09262 (2024)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Liu, M., Wang, J., Qian, X., Li, H.: Listenformer: Responsive listening head generation with non-autoregressive transformers. In: ACM MM. pp. 7094–7103 (2024)
26. Liu, X., Guo, Y., Zhen, C., Li, T., Ao, Y., Yan, P.: Customlistener: Text-guided responsive interaction for user-friendly listening head generation. In: CVPR. pp. 2415–2424 (2024)
27. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
28. Luo, C., Wang, J., Li, B., Song, S., Ghanem, B.: Omniresponse: Online multimodal conversational response generation in dyadic interactions. In: NeurIPS (2025)

29. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: CVPR. pp. 20395–20405 (2022)
30. Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.: From audio to photoreal embodiment: Synthesizing humans in conversations. In: CVPR. pp. 1001–1010 (2024)
31. Park, S., Kim, C., Rha, H., Kim, M., Hong, J., Yeo, J., Ro, Y.: Let's go real talk: Spoken dialogue model for face-to-face conversation. In: ACL (2024)
32. Peng, Z., Fan, Y., Wu, H., Wang, X., Liu, H., He, J., Fan, Z.: Dualtalk: Dual-speaker interaction for 3d talking head conversations. In: CVPR. pp. 21055–21064 (2025)
33. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In: ICCV. pp. 20687–20697 (2023)
34. Pratticó, F.G., Shabkhoslati, J.A., Shaghaghi, N., Lamberti, F.: Bot undercover: on the use of conversational agents to stimulate teacher-students interaction in remote learning. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 277–282. IEEE (2022)
35. Siniukov, M., Chang, D., Tran, M., Gong, H., Chaubey, A., Soleymani, M.: DiTaiListener: Controllable high fidelity listener video generation with diffusion. In: ICCV. pp. 11991–12001 (2025)
36. Sun, Z., Peng, Z., Ma, Y., Chen, Y., Zhou, Z., Zhou, Z., Zhang, G., Zhang, Y., Zhou, Y., Lu, Q., et al.: Streamavatar: Streaming diffusion models for real-time interactive human avatars. arXiv preprint arXiv:2512.22065 (2025)
37. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In: ECCV. pp. 700–717. Springer (2020)
38. Wang, R., Ma, C., Li, G., Xu, H., Li, Y., Wang, Z.: You think, you act: The new task of arbitrary text to motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12012–12022 (2025)
39. Xie, Y., Gu, T., Li, Z., Zhang, C., Song, G., Zhao, X., Liang, C., Jiang, J., Xu, H., Luo, L.: X-streamer: Unified human world modeling with audiovisual interaction. arXiv preprint arXiv:2509.21574 (2025)
40. Xu, J., Guo, Z., Hu, H., Chu, Y., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
41. Xue, H., Fan, Y., Wang, X., Wu, Z.: Echo: Enhancing conversational behavior generation via hierarchical semantic comprehension with large language models. In: SIGGRAPH Asia. pp. 1–9 (2025)
42. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: CVPR. pp. 1087–1096 (2019)
43. Yang, Y., Cen, Z., Peng, S., Chen, X., Deng, Y., Zhu, X., Jia, F., Zhou, X., Bao, H.: StreamingTalker: Audio-driven 3d facial animation with autoregressive diffusion model. In: AAAI (2025)
44. Zhang, Y., Zhong, Z., Liu, M., Chen, Z., Wu, B., Zeng, Y., Zhan, C., He, Y., Huang, J., Zhou, W.: Musetalk: Real-time high-fidelity video dubbing via spatio-temporal sampling. arXiv preprint arXiv:2410.10122 (2024)
45. Zhang, Z., Zhou, Y., Yao, H., Ao, T., Zhan, X., Liu, L.: Social agent: Mastering dyadic nonverbal behavior generation via conversational llm agents. In: ACM SIGGRAPH Asia. pp. 1–12 (2025)

46. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: CVPR. pp. 3661–3670 (2021)
47. Zhao, X., Dai, J., Zhou, F., Wang, H., Song, Z., Hao, A., Qin, H., Gao, Y.: Emoposeface: Head pose aware speech-driven 3d emotional facial animation using latent diffusion. IEEE TVCG (2026)
48. Zhong, Z., Ji, Y., Kong, Z., Liu, Y., Wang, J., Feng, J., Liu, L., Wang, X., Li, Y., She, Y., et al.: Anytalker: Scaling multi-person talking video generation with interactivity refinement. arXiv preprint arXiv:2511.23475 (2025)
49. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T.: Interactive conversational head generation. TPAMI (2025)
50. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., Mei, T.: Responsive listening head generation: a benchmark dataset and baseline. In: ECCV. Springer (2022)
51. Zhu, Y., Zhang, L., Rong, Z., Hu, T., Liang, S., Ge, Z.: INFP: Audio-driven interactive head generation in dyadic conversations. In: CVPR. pp. 10667–10677 (2025)