

A First Exploration of Neuromorphic OT-CFM for Multi-Speaker VSR

Lin Chen¹, Jingping Fang¹, Hairui Liu¹, Chenyang Xu²,
Junhao Chen³, Xiaorui Li¹, Weidong Cai⁴, and Xiaoming Chen^{✉1}

¹ Beijing Technology and Business University, China

² Xidian University, China

³ Tsinghua University, China

⁴ The University of Sydney, Australia

Corresponding Author: Xiaoming Chen (xiaoming.chen@btbu.edu.cn)

Abstract. Visual Speech Recognition (VSR) tasks in complex multi-speaker scenarios are severely hindered by rapid head motions, occlusions, and subtle lip articulations. Traditional RGB-based methods struggle here due to low rates and motion blur of frames. To overcome these, we propose **LipsFlow**, a neuromorphic-inspired VSR framework that converts RGB videos into high-temporal-resolution event streams. For multi-speaker, we employ ByteTrack tracking and TalkNet active speaker detection to temporally segment scenes into single-speaker clips, enabling focused per-speaker analysis. By explicitly capturing microsecond-level articulatory dynamics via learnable event-based representations, LipsFlow achieves inherent robustness against visual degradation. To efficiently model these dense event-based features and adapt to speaker-specific articulatory patterns, we introduce Optimal Transport Conditional Flow Matching (**OT-CFM**). It enforces deterministic, straight-line trajectory generation in a semantic latent space, slashing inference latency to just two Ordinary Differential Equation (ODE) steps. Furthermore, we design a Dual-Level Semantic Supervision mechanism combining token-level BERT weight tying and sentence-level priors to resolve homophone ambiguities. Validated on competitive benchmarks, LipsFlow achieves a state-of-the-art **WER of 22.3% at 240 ms latency**, establishing a highly robust and efficient paradigm for event-based VSR.

Keywords: VSR · Multi-Speaker · Event Streams · OT-CFM

1 Introduction

VSR advances have driven the evolution of multimodal interaction, bridging the sensory gap when traditional acoustic signals are compromised. However, in complex multiple-speaker scenarios, establishing robust visual models independent of auditory cues is significant for sustained communication [22, 37, 46, 56]. While

[✉]Corresponding author.

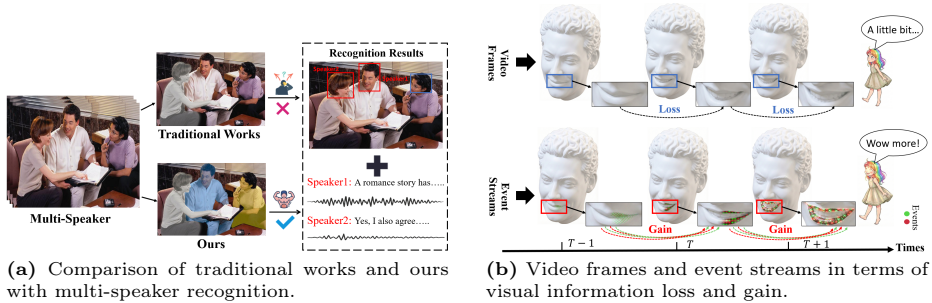


Fig. 1: Our work simultaneously focuses on the following two challenges: **(a)** Traditional VSR methods are only applicable to single-speaker scenarios, while we are able to adapt to multi-speaker scenarios; **(b)** Compared with video frames, event streams provide more subtle information gain for lip reading, thus compensating for the loss of fine-grained motion between consecutive frames.

current mainstream research [3, 22, 23, 37, 56] typically relies on standard active pixel sensors, the inherent limitations of frame-based mechanisms, specifically low fixed frame rates and motion blur, result in the loss of transient visual cues, such as the rapid lip closures of plosives (e.g., /p/ and /b/) that occur within milliseconds [15]. To mitigate these sensing limitations, recent works [3, 47] have introduced neuromorphic event streams to the VSR task, demonstrating the potential of asynchronous event streams to capture microsecond-level lip dynamics [47]. Furthermore, Arriandiga *et al.* [1] have explored event-driven cameras for target speaker enhancement in complex multi-speaker scenarios.

Despite these advancements, current VSR paradigms with event streams [1, 3, 47] face substantial bottlenecks in both representation and generation. First, in terms of vision representation, traditional methods [5, 35] typically aggregate asynchronous events into dense frames, feeding them into standard backbones designed for synchronous data. This approach often fails to disentangle fine-grained phonetic movements from non-phonetic oral noise (e.g., chewing or jitter). Second, the text sequence generation aspect remains a critical efficiency limitation. Dominant approaches rely either on Auto-Regressive (AR) [37], which suffer from prohibitive token-by-token latency, or on standard diffusion models [35], which require extensive stochastic sampling steps to converge, thus affecting text decoding efficiency. Finally, these feed-forward architectures often lack the iterative refinement capabilities needed to retroactively resolve visual ambiguities (e.g., distinguishing “I love you baby” from “Elephant juice bat”).

To overcome these limitations, we simultaneously tackle the challenges of *Multi-Speaker* and *Frame Loss*, as illustrated in Fig. 1. Reconstructing transient phonetic micro-dynamics and resolving the inherent visual ambiguity of lip reading naturally formulate VSR as a complex sequence generation task. While inspired by the powerful diffusion modeling [7, 33–35, 38], we explore its potential for VSR applications. We further introduce OT-CFM decoder based on recent research of Flow Matching [21, 30, 31, 33, 34] to the VSR domain for the first time.

By extending this technique to multi-speaker scenarios and explicitly learning straight trajectories that minimize transport costs, our method achieves a superior trade-off between recognition fidelity and inference efficiency, significantly accelerating model speed without compromising accuracy.

In this work, we propose **LipsFlow**, an end-to-end neuromorphic event-based VSR framework explicitly tailored for complex multi-speaker scenarios. Our method employs a specially designed Event Representation Module for lip reading within multi-speaker environments, explicitly separating fine-grained speech dynamics from non-speech noise, such as chewing or temporal visual occlusion. To resolve homophene ambiguities and ensure global linguistic coherence, we establish a Dual-Level Semantic Supervision strategy leveraging BERT weight tying and Sentence-BERT priors. By combining this strategy with OT-CFM, we explicitly model the generation of continuous, optimal transmission paths for VSR, thereby achieving faster and more accurate training and inference, effectively addressing the fundamental limitations inherent in purely visual methods.

2 Related Works

Visual Speech Recognition Paradigms. VSR has become indispensable for robust speech decoding in acoustically degraded environments [37, 56]. Early end-to-end pioneers [2] leveraged 3D-CNNs and Connectionist Temporal Classification (CTC) loss to bypass complex frame-level alignment. Recent advancements [37, 46] have pushed performance boundaries through massive self-supervised pre-training, which model long-range dependencies via Conformer architectures. Furthermore, state-of-the-art (SOTA) approaches [5, 23, 45] have attempted to enhance robustness via synthetic visual data generation, large language model priors or multi-modal contrastive learning. However, above these systems face an insurmountable physical bottleneck: reliance on standard RGB cameras. Operating at 25–30 FPS, the camera suffers from severe motion blur and inevitably undersamples rapid articulatory micro-dynamics.

Flow Matching Modeling. To address the text mapping in VSR caused by homophenes, generative modeling has evolved from AR sequence prediction to Non-Autoregressive (NAR) backbones. While diffusion-based models [34, 35, 40] improve generation fidelity by iteratively refining noise, they necessitate 50–1000 sampling steps, incurring prohibitive latency and computational costs that hinder real-time deployment. Conditional Flow Matching (CFM) [31] has recently emerged as a superior generative backbone. Compared with the diffusion, which models curved stochastic trajectories, CFM with Optimal Transport (OT) objectives regresses straight-line probability paths [33], enabling the ODE solver [6] to traverse from prior noise to target data in 2-10 NFE steps. While OT-CFM [30] has revolutionized speech generation works [21, 32, 38, 43], its potential in VSR remains unexplored, particularly in multi-modal contrastive learning [7]. By combining efficient OT-CFM [30] generation with Dual-Level Semantic Supervision [42], our approach not only resolves semantic ambiguities inherent to lip reading but also achieves real-time inference speed.

Event-based Representations. Neuromorphic event cameras offer an inspiration for VSR by capturing asynchronous brightness changes with microsecond-level temporal resolution, high dynamic range (> 120 dB), and immunity to motion blur [15]. These properties have driven success in high-speed tasks such as optical flow estimation, object tracking, and detection using recurrent transformers [17]. Despite these sensory advantages, considering event cameras in VSR presents unique challenges due to the sparse nature of events in textureless lip reading and the difficulty of representation learning. Existing event-based VSR methods [10, 47, 54] often aggregate asynchronous events into dense representations like Voxel Grids or Event Stacks to utilize standard CNN backbones. However, these methods typically struggle in unconstrained multi-speaker environments. Generic event-stream encoders fail to disentangle target phonetic movements from non-communicative oral activities (e.g., chewing, yawning) or background noise, as both generate dense event streams. Furthermore, unlike RGB-based VSR which benefits from massive datasets [37], event-based VSR suffers from data scarcity. While simulation techniques Video-to-Events (V2E) [16] exist, there is a lack of frameworks that effectively bridge the gap between large-scale RGB pre-training and high-frequency event sensing.

3 Proposed Method

3.1 Overview

We introduce **LipsFlow** framework in Fig. 2 for handling multi-speaker scenarios in VSR under various complex conditions. In order to tackle multi-speaker scenarios in batches, we employ the ByteTrack [55] front-end to physically isolate each speaker for focused processing. Rapid articulatory dynamics are captured via a **Learnable Event-based Representation** (see Sec. 3.2) module that extracts high-frequency features from RGB videos. These features condition our **Speaker-Conditioned OT-CFM** (see Sec. 3.3), which efficiently generates semantic tokens via a straight-line probability flow. We resolve homophone ambiguities through a **Prior-Guided Semantic Decoder** (see Sec. 3.4), which leverages linguistic priors to enforce global coherence and contextual consistency.

3.2 Learnable Event-based Representation

Hierarchical Frame Interpolation. Reconstructing high-frequency lip dynamics from RGB videos is an ill-posed problem. Direct $20\times$ upsampling often fails as large non-linear displacements between frames exceed the effective receptive field of multi-speaker optical streams, leading to severe ghosting artifacts. To address this, we devise a hierarchical strategy that decomposes the task into a cascade of manageable stages: $2\times \rightarrow 2\times \rightarrow 5\times$. At each level, we leverage a lightweight variant of RIFE [26] to estimate bidirectional optical streams. By progressively subdividing the temporal intervals, this strategy reduces the motion magnitude between adjacent frames, ensuring robust streams estimation for rapid articulatory motions via the lightweight backbone design.

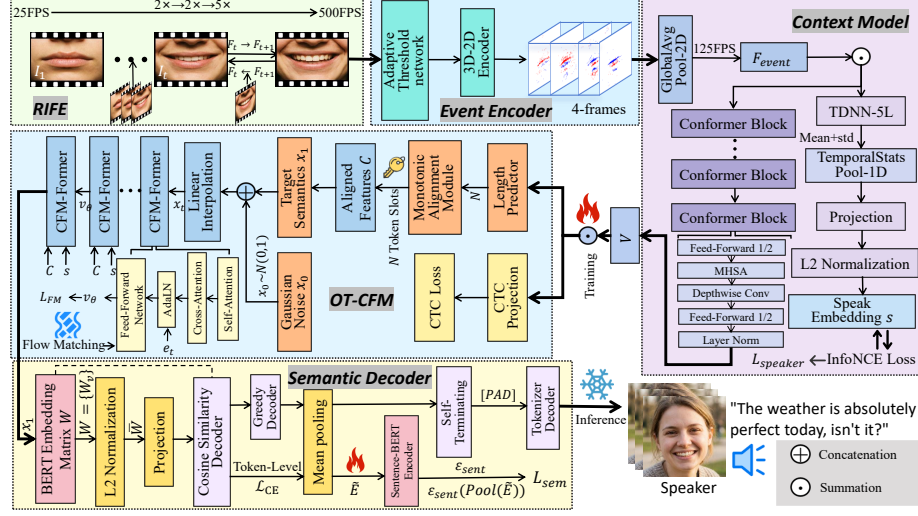


Fig. 2: Overview of LipsFlow Framework. Event-based visual features are extracted via hierarchical interpolation and Conformer encoding, then conditioned on speaker embeddings through OT-CFM to generate semantic token representations. The BERT [12] weight tying is utilized alongside multi-level supervision signals during training (token-level $\mathcal{L}_{CE}$, sentence-level $\mathcal{L}_{sem}$, and auxiliary CTC loss $\mathcal{L}_{CTC}$ for alignment), ensuring both local precision and global semantic coherence.

Adaptive Event Streams Generation. We transform the upsampled frames into high-dynamic-range event streams using a differentiable quantizer [14]. We introduce an adaptive thresholding network that leverages a lightweight CNN to regress a dense threshold map Θ . An event is triggered at coordinate $\mathbf{x}$ if the logarithmic temporal residual exceeds the local threshold $\theta(\mathbf{x})$:

$$|\Delta L(\mathbf{x})| = |\log I_{t+\Delta t}(\mathbf{x}) - \log I_t(\mathbf{x})| > \theta(\mathbf{x}). \quad (1)$$

The soft relaxation strategy [27] is employed during training to ensure end-to-end differentiability, enabling the network to selectively amplify articulatory micro-kinematics to better capture subtle and rapid information about lip movements.

Event Encoding and Context Modeling. To encode the high FPS of event streams, a hybrid 3D-2D encoder (3D Stem + 2D ResNet [24]) aggregates local motion, yielding a feature sequence $F_{event} \in \mathbb{R}^{T' \times 512}$ at 125 FPS. Since F_{event} captures only local cues, we feed it into the 12-layer Conformer [20] encoder to model global temporal dependencies. Synergizing Depthwise Convolution [25] and Multi-Head Self-Attention (MHSA) [49], the Conformer produces context-aware visual features $V \in \mathbb{R}^{T' \times 512}$. This hierarchical Visual Encoder design allows LipsFlow to effectively model high-frequency transient dynamics of rapid lip movements while maintaining the long-range temporal dependencies.

3.3 Speaker-Conditioned OT-CFM

Length Prediction and Alignment. A critical challenge in NAR speech generation is determining the target sequence length N during inference. Instead of relying solely on implicit attention, we employ a Hybrid Length Predictor [43] (combining regression and classification heads) to estimate the number of semantic tokens from the visual features V . Subsequently, a monotonic alignment module projects the high-frequency visual streams (125 Hz) onto the predicted N token slots, producing aligned visual conditions $C \in \mathbb{R}^{N \times 512}$. While the CTC [19] head is used as an auxiliary loss during training to enforce alignment constraints, the explicit length predictor governs the generation structure during inference.

Speaker Identity Encoding and Disentanglement. To disentangle target speech dynamics from multi-speaker interferences, we extract a global speaker embedding $s \in \mathbb{R}^{256}$ using a Time-Delay Neural Network (TDNN) [50] on F_{event} . A temporal statistics pooling layer aggregates identity-specific attributes across time. Crucially, we utilize an InfoNCE [39] contrastive loss to enforce neural disentanglement: by clustering same-speaker embeddings, s functions as an active semantic anchor that filters out identity-irrelevant visual noise (e.g., background motion), transcending simple physical cropping.

Optimal Transport Flow Matching. Rather than regressing discrete tokens, we reformulate generation as learning a continuous probability flow ODE [6] within a pre-trained semantic latent space. Building on CFM [31], we model the transformation from a Gaussian prior $x_0 \sim \mathcal{N}(0, 1)$ to BERT-encoded [12] target semantics x_1 . To guarantee real-time efficiency, we adopt the Optimal Transport (OT) path [7, 30, 33], simplifying our training objective to regressing a vector field that matches the straight-line displacement between these distributions. Moreover, we directly condition this vector field on our aligned visual features V and speaker embedding s , ensuring the sequence generation is precisely driven by target lip dynamics and identity. Unlike CTC-based methods [2, 19, 37], OT-CFM models the distribution of phonetic trajectories, enabling the decoder to resolve visual ambiguities through iterative refinement in the latent space.

Dual-Conditioned Architecture. We parameterize the velocity estimator v_θ using a 16-layer CFM-Former designed to disentangle linguistic content from speaker identity. At each layer, aligned visual conditions C are injected via Cross-Attention [49], where the evolving state x_t serves as queries and C provides keys and values. Simultaneously, speaker identity is strictly integrated through Adaptive Layer Normalization (AdaLN) [40]:

$$\text{AdaLN}(h, e_t, s) = \gamma(e_t, s) \cdot \text{LN}(h) + \beta(e_t, s), \quad (2)$$

where the affine parameters γ, β are dynamically modulated by the time embedding e_t and speaker embedding s . This design effectively decouples content generation from identity modulation: Cross-Attention enforces semantic alignment with lip movements, while AdaLN [40] injects global speaker traits. Finally, leveraging the geometrically straightened flow path [33] modeled by this architecture, we conduct inference by solving the probability flow ODE [6] via a

deterministic 2-step Euler solver ($\Delta t = 0.5$). This process recovers the target semantic representations $x_1 \in \mathbb{R}^{N \times 512}$.

3.4 Prior-Guided Semantic Decoder

Cross-Modal Weight Tying with BERT. To bridge the modality gap between continuous visual kinematics and discrete linguistic semantics, we avoid learning an isolated projection head. Instead, we map the recovered features x_1 into the pre-trained BERT [12] embedding space W , leveraging it as a strong semantic anchor. Specifically, a learnable projection $\phi(\cdot)$ transforms x_1 to align with the semantic manifold of W . We then compute the probability distribution over the vocabulary, $\hat{\mathbf{y}}$, via temperature-scaled cosine similarity:

$$\hat{\mathbf{y}} = \text{Softmax} \left(\frac{\phi(x_1) \bar{W}^\top}{\tau \cdot \|\phi(x_1)\|_2} \right), \quad (3)$$

where $\bar{W}$ is the normalized embedding matrix and τ is a learnable scalar. This formulation enforces directional alignment within the hypersphere, transferring rich linguistic priors to resolve visual ambiguities without massive parameters.

Dual-Level Semantic Supervision. Although $\mathcal{L}_{\text{CE}}$ ensures local token precision, it is inherently myopic and suffers from lexical rigidity, failing to resolve visual ambiguities arising from homophenes. Therefore, we augment the training objective with a global semantic consistency constraint via a frozen Sentence-BERT [42] encoder $\mathcal{E}_{\text{sent}}$. To preserve end-to-end differentiability, we bypass the non-differentiable discrete decoding step (e.g., argmax). Instead, we introduce soft-weighted [27] token embeddings as the expectation over the predicted distribution. Specifically, the continuous embedding at step i is calculated as the weighted sum of all fixed BERT [12] vocabulary embeddings, where each token’s embedding is scaled by its corresponding predicted probability $\hat{y}_{i,v}$ at that step.

The resulting sequence $\tilde{\mathbf{E}} = [\tilde{\mathbf{e}}_1, \dots, \tilde{\mathbf{e}}_N]$ is aggregated via mean pooling and fed to Sentence-BERT to minimize the semantic discrepancy:

$$\mathcal{L}_{\text{sem}} = \left\| \mathcal{E}_{\text{sent}}(\text{Pool}(\tilde{\mathbf{E}})) - \mathcal{E}_{\text{sent}}(\mathbf{y}) \right\|_2^2. \quad (4)$$

This formulation ensures gradient flow throughout the entire pipeline, enabling the Visual Encoder (see Sec. 3.2) to learn discriminative features for homophone disambiguation. The synergistic optimization of structural flow ($\mathcal{L}_{\text{FM}}$), lexical precision ($\mathcal{L}_{\text{CE}}$), and semantic consistency ($\mathcal{L}_{\text{sem}}$) empowers our framework to robustly disambiguate homophenes.

Adaptive Termination. To enable the fixed-dimensional flow model to generate variable-length sequences, we introduce a Self-Terminating Decoder. We explicitly train the model to recognize sequence boundaries by stochastically appending padding tokens ([PAD]) to the training targets [18]. This compels the model to internalize the termination signal within the semantic manifold. The final transcription is truncated at the first termination token, enabling flexible sequence lengths.

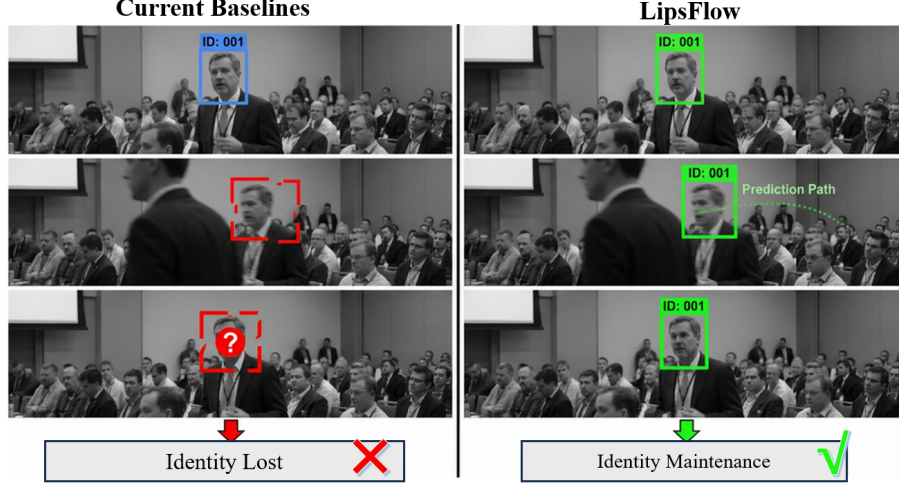


Fig. 3: Robustness of Identity Tracking in Multi-Speaker. Existing baselines (**Left**) suffer from identity loss due to dynamic disturbances (**red** dashed boxes). In contrast, our method LipsFlow (**Right**) effectively maintains speaker identity continuity through its predicted path mechanism, ensuring stable identity tracking even in challenging environments (**green** solid boxes).

4 Multi-Stage Data Processing Pipeline

While Sec. 3 details the LipsFlow architecture, its performance hinges on high-quality input. We establish a powerful **Multi-Stage Data Process Pipeline** designed to isolate target speakers in complex, multi-speaker environments. The pipeline implements a rigorous curation protocol to address three critical challenges: maintaining identity consistency under occlusion, extracting stable facial features, and filtering non-speech visual noise with specific annotation labels.

4.1 Identity Consistency Maintenance

We operate on the source RGB videos and build our data processing pipeline upon the robust ByteTrack [55]. Specifically, we leverage its hierarchical recovery strategy supported by Kalman filtering [4] and IoU-based association [13] to rescue low-confidence detections from **temporary occlusion** or **pose variations** (see Fig. 3). Beyond applying standard tracking, we propose a tailored spatiotemporal trajectory refinement protocol optimized for event-based vision. We enforce a customized temporal continuity check [41] to actively merge fragmented segments and suppress transient false alarms in dynamic conversational scenarios. Significantly, we design this specialized pipeline not only to guarantee strictly aligned identity supervision without identity loss, but also to act as a dedicated physical noise-suppressor.

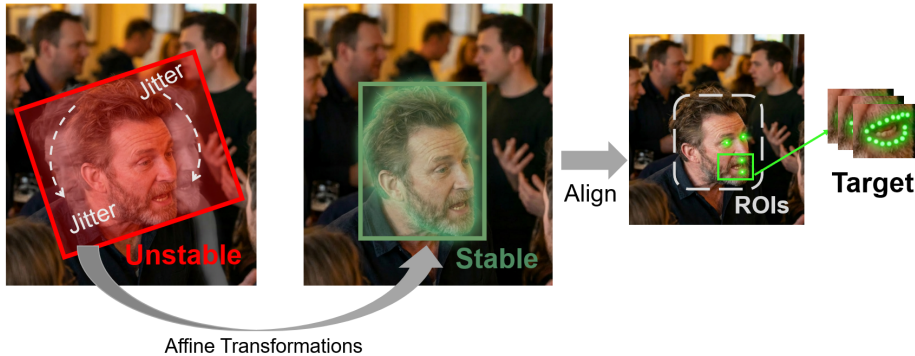


Fig. 4: Stabilization and Extraction of Features. The video frames (**Unstable**) typically contain significant jitter, which leads to severe noise during events generation. By applying the Affine Transformation based on facial feature points, we align the ROIs to obtain a stable target region (**Stable**). This process ensures that subsequently generated events reflect only lip dynamics, not head movements.

4.2 ROIs Extraction and Stabilization

We engineer a cascaded system for precise Lip Region of Interest (ROIs) extraction (see Fig. 4). Following identity tracking, we deploy RetinaFace [11] for high-precision face detection, succeeded by a Dlib [29] regression tree model to localize 68 facial landmarks. We specifically extract the 20 lip landmarks (indices 48–67) incorporating a 20% contextual padding. To eliminate rigid **head jitter** that corrupts event generation, we apply Affine Transformations [37] to align the ROIs based on stable eye and nose anchors. To prevent false event triggers caused by artificial zero-padding boundaries introduced during this affine alignment, we apply a strict central crop and masking strategy. Finally, we normalize the purified crops to 128×64 using gradient-preserving interpolation, establishing a consistent, noise-free spatial reference tailored for downstream neuromorphic event streams generation with Sec. 3.2.

4.3 Annotation Protocol

We establish an annotation protocol to distinguish valid speech sources from background noise. Following the common practice [44], we define a three-tier vocabulary: (1) *Silence*; (2) *Oral Motion* (e.g., only chewing); and (3) *Active*. Fundamentally, this explicit state decoupling is an indispensable requirement for event-based VSR. Because neuromorphic sensors respond purely to temporal kinematics, state (2) triggers dense event streams that are visually indistinguishable from actual speech, which would otherwise induce severe semantic hallucinations in generative models. To resolve these visual-audio discrepancies, we leverage synchronized audio waveforms as a semantic ground-truth anchor to precisely calibrate the temporal boundaries of (3).

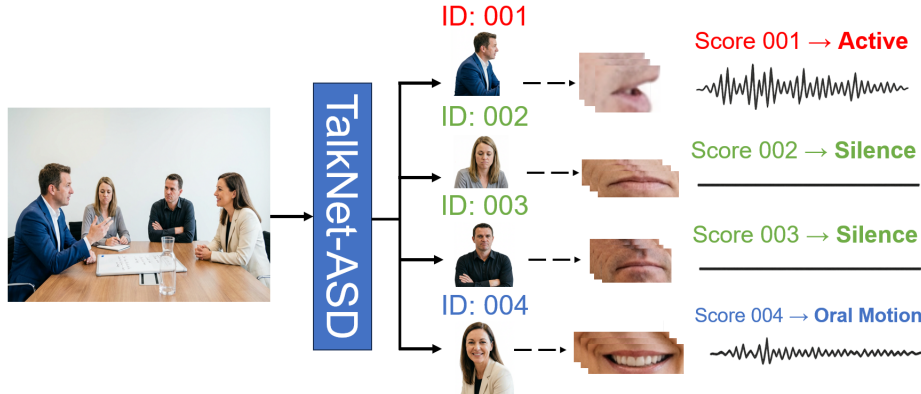


Fig. 5: Multi-Speaker Interference with Active Speaker Detection. Our system assigns a specific annotation state to the tracked identity (see Sec. 4.3), enabling the model to pinpoint the primary speaker and ignore visual interference from passive listeners. Note, we retain only identity-audio pairs exceeding a strict confidence threshold.

4.4 Multi-Speaker Active Detection

While ByteTrack [55] maintains visual identity, the association of these identities with overlapping audio streams in multi-speaker scenarios presents a distinct challenge. We resolve this **identity-to-audio** mapping problem by leveraging TalkNet-ASD [48] to improve the accuracy of audio-visual consistency in matching identity. As illustrated in Fig. 5, this strategy computes correspondence scores between each tracked identity and temporal audio segments by evaluating the correlation between macroscopic lip motions and mel-spectrogram features.

5 Experiments and Results

5.1 Implementation Details

Datasets. We evaluate LipsFlow and baselines on two complementary benchmarks: **DVS-Lip** [47] and **AVA** [44]. The former constitutes the first large-scale corpus with real-world event streams, encompassing not only subtle lip micro-movements but also challenging vocabulary. Note, we utilize visual information (voxel or event streams) directly from DVS-Lip. Conversely, the AVA introduces uncontrolled, natural scene complexity multi-speaker scenarios, which can convert frames into event streams via our Data Process Pipeline (see Sec. 4). We perform three classification annotations on the AVA with video keyword retrieval and human screening, categorizing it into *Rapid Motion*, *Low Light*, and *Severe Occlusion*, to validate the model’s adaptability for complex scenarios.

Data Augmentation. We further validate the adaptability of LipsFlow and baselines to the environment through a standard data augmentation mechanism. This process includes four probabilistic transformations: **MixUp** (probability 0.3) [53], which promotes linear behavior between categories through

Table 1: Comparison Benchmarking Results on **DVS-Lip**. Note that we utilize the event streams provided by this dataset directly as input, rather than simulations. Metrics: *WER* / *VER* (Word [2]/Viseme [47] Error Rate, %, $\downarrow$), *RTF* (Real-Time Factor [43], s), *Latency* (inference in ms), and *Params* (Model Parameters, M). **Bold** and underlined values denote the best and second-best performance. While Llama-AVSR [5] achieves competitive accuracy, our method balances high fidelity with superior efficiency. The principles for reproduction all follow those outlined on the official website.

Method	Modality	Backbone	Accuracy ($\downarrow$)		Efficiency ($\downarrow$)		
			WER	VER	RTF	Latency	Params
<i>Video-based: Standard Architectures</i>							
Auto-AVSR [37]	RGB	Conformer	28.4	26.1	0.45	580	256.4
VATLM [56]	RGB	Transformer	26.9	25.3	0.52	650	310.8
RAVE _n [22]	RGB	Conformer	31.2	29.8	0.15	200	55.2
SynthVSR [35]	RGB	Conformer	25.1	23.5	4.80	6000	185.6
<i>Video-based: Recent SOTA</i>							
CMAVSR [45]	RGB	Transformer	27.5	25.8	0.46	590	260.1
LipGen [23]	RGB	ResNet-18	26.1	24.6	0.55	680	235.5
Llama-AVSR [5]	RGB	LLM	<u>23.9</u>	<u>21.5</u>	2.10	2500	>1000
<i>Event-based Methods</i>							
MSTP [47]	Event	ResNet-18	30.5	28.2	0.22	280	22.5
SNN-Lip [3]	Event	SNNs	34.1	32.4	<u>0.18</u>	230	12.1
LipsFlow (Ours)	Event	CFM-Former	22.3	19.8	<u>0.18</u>	<u>240</u>	<u>45.8</u>

Beta distribution interpolation; **CutMix** (probability 0.3) [52], which replaces fragments with consecutive 25% to 50% segments, forcing the network to learn temporal continuity rather than memorizing local patterns; **Occlusion** (probability 0.2), which simulates sensor malfunctions or frame jitter by masking subsequences [51]; and **Identity Mapping** (probability 0.2), used to maintain the fidelity of the distribution. This comprehensive strategy effectively simulates imperfections in real-world videos, thus ensuring the model’s robustness [37].

Training. We orchestrate the training process on 4 NVIDIA A100 (80GB) GPUs via the PyTorch Distributed Data Parallel framework. We strictly follow the official train/test protocols of the DVS-Lip [47] and AVA [44] datasets to ensure a fair comparison with baselines. Furthermore, we leverage BF16 mixed precision [28] and TF32 acceleration [8] under a global batch size of 64 (32 sequences per GPU) to balance numerical fidelity with computational throughput. We design a rigorous two-stage curriculum to decouple feature extraction from generation: initially, we *freeze* the visual front-end for the first 10 epochs ($\approx 15\text{k}$ iterations, $lr = 5 \times 10^{-4}$) to align the generative manifold without destabilizing the pretrained representations; subsequently, we *unfreeze* the entire architecture for holistic end-to-end fine-tuning for another 20 epochs ($\approx 30\text{k}$ iterations, $lr = 1 \times 10^{-4}$). We drive convergence using the AdamW [36] optimizer ($\beta_1 = 0.9, \beta_2 = 0.995$) and safeguard training stability via gradient norm clipping (1.0). The learning rate is modulated by a cosine annealing scheduler with 2,000 warmup steps, concluding the entire 30 hours training regimen after approximately 45k total iterations.

Table 2: Robustness Benchmarking Results on **AVA** with WER (%). *Clean Set* aligns with the main results in Tab. 1. Values in parentheses denote relative degradation ($\uparrow$ performance drop). Values in red/orange/green font indicate severe/moderate/slight performance degradation, respectively, with smaller values representing better results. Note, we use the adaptive threshold network (see Sec. 3.2) and data processing pipeline (see Sec. 4) to generate the event streams.

Method	Clean Set	Rapid Motion	Low Light	Severe Occlusion
<i>Video-based</i>				
Auto-AVSR [37]	28.4	45.2 ($\uparrow 16.8$)	52.1 ($\uparrow 23.7$)	48.9 ($\uparrow 20.5$)
LipGen [23]	26.1	41.8 ($\uparrow 15.7$)	49.5 ($\uparrow 23.4$)	44.2 ($\uparrow 18.1$)
SynthVSR [35]	25.1	39.4 ($\uparrow 14.3$)	47.1 ($\uparrow 22.0$)	38.6 ($\uparrow 13.5$)
Llama-AVSR [5]	23.9	36.5 ($\uparrow 12.6$)	44.3 ($\uparrow 20.4$)	35.1 ($\uparrow 11.2$)
<i>Event-based</i>				
MSTP [47]	30.5	33.1 ($\uparrow 2.6$)	31.5 ($\uparrow 1.0$)	42.8 ($\uparrow 12.3$)
LipsFlow (Ours)	22.3	24.1 ($\uparrow 1.8$)	22.9 ($\uparrow 0.6$)	26.5 ($\uparrow 4.2$)

Baseline Systems. We compare LipsFlow within selective landscapes across two modalities (RGB and Events). Regarding **Video-based** methods, we benchmark against the established *Auto-AVSR* [37] and the pre-trained *VATLM* [56] with Masked Prediction architecture, while incorporating current SOTA research: the robust *RAVEN* [22], the contrastive *CMAVSR* [45], the generation-enhanced *LipGen* [23], and the large-model-augmented *Llama-AVSR* [5]. Furthermore, we juxtapose our generative efficiency with the Conformer-based *SynthVSR* [35], highlighting our OT-CFM paradigm. We scrutinize our performance against specialized **Event-based** methods, specifically the *MSTP* [47] and the parameter-efficient *SNN-Lip* [3], to maintain the comprehensiveness of the RGB and Events baselines comparison. Note, the video-based methods input only uses RGB video information, while the event-based methods input requires additional generated neuromorphic event streams for a fair comparison.

5.2 Comparative Benchmarking Results

As quantified in Tab. 1, we demonstrate that LipsFlow establishes a new performance ceiling on the multi-speaker event-based benchmark DVS-Lip for VSR, securing a WER of **22.3%** and a VER of **19.8%**. Notably, our method outperforms the sophisticated video-based baselines, surpassing the generation-enhanced LipGen by 3.8% and even edging out the large-scale Llama-AVSR by 1.6%. This result highlights a critical insight: our method captures transient phonetic micro-dynamics, such as rapid plosives, more effectively than simply scaling up model parameters or utilizing massive pre-trained language backbones on standard RGB video. Simultaneously, we surpass the leading event-based competitor (MSTP) by 8.2%, validating the architectural superiority of our CFM-Former backbone over traditional ResNet-18 paradigms in modeling complex temporal dependencies. While our WER (22.3%) is highly competitive with the Conformer-based SynthVSR (22.5%), we drastically slash the RTF from

Table 3: Ablation Study on Architecture Variants. *Fusion Strategy* contrasts naive channel-wise concatenation (*Concat*) with our dual-branch temporal alignment via Cross-Attention (*Cross-Attn*). Paradigm benchmarks standard Auto-Regressive (*AR*) Transformers and Denoising Diffusion (*DDPM*) against our OT-CFM decoder.

Variants	Input		Fusion	Decoder	Accuracy ($\downarrow$)		
	RGB	Event	Strategy	Paradigm	WER	VER	Params
(a) Visual-Only	✓	✗	N/A	AR	27.8	25.9	42.1
(b) Naive Fusion	✓	✓	Concat	AR	25.9	24.1	44.5
(c) Neuromorphic Branch	✓	✓	Cross-Attn	AR	24.5	22.8	46.2
(d) Diffusion Variant	✓	✓	Cross-Attn	Diffusion	22.5	20.1	185.6
(e) LipsFlow (Ours)	✓	✓	Cross-Attn	OT-CFM	22.3	19.8	45.8

4.80 to 0.18, a staggering $26\times$ acceleration. This speedup stems directly from our introduced OT path, which enables high speed inference in just 2 Number of Function Evaluations (NFE) compared to prior methods.

5.3 Environmental Resilience and Degradation

As quantified in Tab. 2, LipsFlow exhibits exceptional resilience against environmental degradation. A pivotal finding is that the limitations of RGB-based sensing persist regardless of algorithmic sophistication. Even the generative model SynthVSR and the large-scale LLM-augmented systems Llama-AVSR suffer catastrophic performance drops. This stems from the fact that pure-pixel sensors are physically incapable of recovering articulatory details lost to motion blur and photon scarcity. In contrast, by shifting the sensing paradigm to neuromorphic event streams, LipsFlow maintains near-negligible degradation (e.g., only $\uparrow 0.6\%$ in Low Light). Furthermore, while Llama-AVSR leverages massive linguistic priors to mitigate Severe Occlusion, our method still outperforms it by a significant margin (26.5% vs. 35.1% WER), demonstrating that neuromorphic event streams representation provides more effective and richer visual information for VSR compared to traditional video-based recognition methods.

5.4 Ablation Study

We conduct a comprehensive ablation study on the **DVS-Lip** [47] to verify the effectiveness of our method components. We specifically ablate our multi-modal fusion, OT-CFM decoder efficiency, and joint optimization objectives.

Impact of Multi-modal Fusion and Architecture. Tab. 3 confirms that the Visual-Only (a) variant achieves a WER of 27.8%, verifying the physical bottleneck of RGB sensors in capturing transient phonemes under rapid articulatory motion. Introducing the Neuromorphic Branch (c) via our Cross-Attention fusion breaks this ceiling to 24.5% by effectively aligning asynchronous event spikes with synchronous video frames, significantly outperforming the channel-wise naive concatenation strategy in (b). Furthermore, replacing standard decoders with

Table 4: Ablation Study on Inference Dynamics. *NFE* vs. *Efficiency* Trade-off. Analyzing sampling steps against fidelity and speed. *Speedup*: Relative to Diffusion.

Variants	NFE	WER↓	Latency↓	RTF↓	Speedup↑
	1	28.5	180	0.09	33.3×
LipsFlow (Ours)	2	22.3	240	0.18	25.0×
	4	22.1	350	0.26	17.1×
	10	22.0	680	0.51	8.8×
<i>Comparison Results:</i>					
Diffusion Variant	50	22.5	6000	4.80	1.0×

Table 5: Ablation Study on Loss Configuration and Training Strategy. *CE* represents the Cross-Entropy loss. We verify the stepwise contribution of optimization objectives and the 2-stage curriculum. *SpkVA*: Speaker Verification Accuracy [9].

Loss Configuration	Stage	WER↓	SpkVA↑
CE-Only	Single	35.8	76.3
OT-CFM	Single	28.2	78.5
+ Speaker Contrastive	Single	26.6	91.7
Full Loss	Single	25.2	92.3
<i>Impact of Training Strategy (with Full Loss):</i>			
Full Loss	Single	25.2	92.3
Full Loss	2-Stage	23.8	93.1
Full Loss + Data Aug	2-Stage	22.3	93.8

the generative LipsFlow (e) yields the final leap; while the Diffusion variant (d) is accurate, it is parameter-heavy (185.6M), whereas our full model achieves the best WER (22.3%) with 4× fewer parameters, demonstrating superior efficiency in modeling complex phonetic distributions within the semantic latent space.

Efficiency of Rectified Flow Trajectories. Tab. 4 shows the critical speed-accuracy trade-off. While single-step generation of NFE suffers from a notable truncation error (28.5% WER) due to the slight residual curvature in the learned vector field, increasing the budget to just NFE=2 yields an optimal performance of 22.3%. This dramatic 6.2% improvement with only one additional integration step empirically validates the core advantage of our OT objective: it geometrically straightens the probability flow so effectively that a simple 2-step Euler solver can traverse the generative manifold with almost zero discretization error. Consequently, we achieve an RTF of 0.18, a massive 25.0× speedup over the curved trajectory of Diffusion (6000ms), thereby ensuring real-time feasibility.

Optimization and Training Strategy. Tab. 5 illustrates the systematic contribution of each learning component. Replacing the discriminative CE-Only (35.8% WER) with our generative OT-CFM drastically lowers WER to 28.2%, proving that modeling **continuous** probability flows **along optimal trajectories** better captures phonetic diversity than simple classification. Additionally, we introduce Speaker Contrastive Learning to decouple features, improving SpkVA to 92.3% by effectively filtering out identity-irrelevant visual disturbances.

Regarding the curriculum, our proposed 2-Stage Strategy (freeze-then-fine-tune) prevents the collapse of pre-trained manifolds during initial convergence, yielding a 1.4% gain. Finally, applying the Data Augmentation (*Data Aug*) in Sec. 5.1 (MixUp, CutMix) secures the performance WER of 22.3%, ensuring robust generalization against unforeseen real-world noise and environmental variability.

6 Conclusion

In this work, we first present **LipsFlow**, a novel framework that fundamentally reimagines VSR by bridging neuromorphic perception with deterministic generative modeling. To overcome the physical limitations of standard cameras and the prohibitive latency of diffusion models, we introduce a paradigm built on two core pillars. We leverage high-frequency event streams to capture transient lip dynamics, completely bypassing the redundancy and motion blur of frame-based video. We employ OT-CFM to learn deterministic, straight-line ODE trajectories, thereby resolving the historical tension between generation quality and inference speed with our data processing pipeline. Furthermore, we address the inherent challenge of homophene ambiguity and non-phonetic oral recognition, which is precisely what traditional methods have failed to solve.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 62577004) and Beijing Natural Science Foundation (No. L2611045).

References

1. Arriandiaga, A., Morrone, G., Pasa, L., Badino, L., Bartolozzi, C.: Audio-visual target speaker enhancement on multi-talker environment using event-driven cameras. arXiv preprint arXiv:1912.02671 (2021)
2. Assael, Y.M., Shillingford, B., Whiteson, S., Freitas, N.D.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)
3. Bulzomi, H., Schweiker, M., Gruel, A., Martinet, J.: End-to-end neuromorphic lipreading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4101–4108. IEEE (2023)
4. Cao, J., Pang, J., Weng, X., Khrodar, R., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9686–9696 (2023)
5. Cappellazzo, U., Kim, M., Chen, H., Ma, P., Petridis, S., Falavigna, D., Brutti, A., Pantic, M.: Large language models are strong audio-visual speech recognition learners. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2025)
6. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural Ordinary Differential Equations. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)

7. Choi, J., Kim, J.H., Li, J., Chung, J.S., Liu, S.: V2sflow: Video-to-speech generation with speech decomposition and rectified flow. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025)
8. Choquette, J., Gandhi, W., Giroux, O., Stam, N., Krashinsky, R.: NVIDIA A100 Tensor Core GPU: Performance and Innovation. *IEEE Micro* **41**(2), 29–35 (2021)
9. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep Speaker Recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). pp. 1086–1090 (2018)
10. Dampfhofer, M., Mesquida, T.: Neuromorphic lip-reading with signed spiking gated recurrent units. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2141–2151 (2024)
11. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 5203–5212 (2020)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL). pp. 4171–4186 (2019)
13. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: StrongSORT: Make DeepSORT Great Again. *IEEE Transactions on Multimedia* **25**, 8725–8737 (2023)
14. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. In: International Conference on Learning Representations (ICLR) (2020)
15. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(1), 154–180 (2022)
16. Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Video to events: Recycling video datasets for event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3586–3595 (2020)
17. Gehrig, M., Scaramuzza, D.: Recurrent vision transformers for object detection with event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13884–13893 (2023)
18. Ghazvininejad, M., Levy, O., Liu, Y., Zettlemoyer, L.: Mask-predict: Parallel decoding of conditional masked language models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6112–6121 (2019)
19. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). pp. 369–376 (2006)
20. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R.: Conformer: Convolution-augmented transformer for speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). pp. 5036–5040 (2020)
21. Guo, Y., Du, C., Ma, Z., Chen, X., Yu, K.: VoiceFlow: Efficient text-to-speech with rectified flow matching. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024)

22. Haliassos, A., Ma, P., Mira, R., Petridis, S., Pantic, M.: Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246* (2022)
23. Hao, B., Zhou, D., et al.: Lipgen: Viseme-guided lip video generation for enhancing visual speech recognition. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2025)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
25. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
26. Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S.: Rife: Real-time intermediate flow estimation for video frame interpolation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 624–642 (2022)
27. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-Softmax. In: *International Conference on Learning Representations (ICLR)* (2017)
28. Kalamkar, D., Mudigere, D., Mellempudi, N., Das, D., Banerjee, K., Avancha, S., Vooturi, D.T., Jammalamadaka, N., Huang, J., Yuen, H., Yang, J., Park, J., Heinecke, A., Georganas, E., Srinivasan, S., Kundu, A., Smelyanskiy, M., Kaul, B., Dubey, P.: A Study of BFLOAT16 for Deep Learning Training. *arXiv preprint arXiv:1905.12322* (2019)
29. King, D.E.: Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* **10**, 1755–1758 (2009)
30. Kornilov, N., Mokrov, P., Gasnikov, A., Korotin, A.: Optimal Flow Matching: Learning Straight Trajectories in Just One Step. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2024)
31. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: *International Conference on Learning Representations (ICLR)* (2023)
32. Liu, A.H., Le, M., Vyas, A., Shi, B., Tjandra, A., Hsu, W.N.: Generative pre-training for speech with flow matching. In: *The Twelfth International Conference on Learning Representations (ICLR)* (2024)
33. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022)
34. Liu, X., Zhang, X., Ma, J., Peng, J., Liu, Q.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In: *International Conference on Learning Representations (ICLR)* (2024)
35. Liu, X., Lakomkin, E., Vougioukas, K., Ma, P., Chen, H., Xie, R., Doulaty, M., Moritz, N., Kolář, J., Petridis, S., Pantic, M., Fuegen, C.: Synthvsr: Scaling up visual speech recognition with synthetic supervision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 18806–18815 (2023)
36. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: *International Conference on Learning Representations (ICLR)* (2019)
37. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-avs: Audio-visual speech recognition with automatic labels. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. p. 1–5 (2023)

38. Mehta, S., Tu, R., Beskow, J., Székely, É., Henter, G.E.: Matcha-tts: A fast tts architecture with conditional flow matching. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024)
39. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
40. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (2023)
41. Qin, Z., Zhou, S., Wang, L., Duan, J., Hua, G., Tang, W.: MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17939–17948 (2023)
42. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP). pp. 3982–3992 (2019)
43. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: FastSpeech 2: Fast and high-quality end-to-end text to speech. In: International Conference on Learning Representations (ICLR) (2021)
44. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., Pantofaru, C.: AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4492–4496. IEEE (2020)
45. Shen, X., Wang, X., Shen, L., Zhang, K., Yu, X.: Cross-view isolated sign language recognition via view synthesis and feature disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20647–20657. IEEE (2025)
46. Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. In: International Conference on Learning Representations (ICLR) (2022)
47. Tan, G., Wang, Y., Han, H., Cao, Y., Wu, F., Zha, Z.J.: Multi-Grained Spatio-Temporal Features Perceived Network for Event-Based Lip-Reading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20062–20071 (2022)
48. Tao, R., Pan, Z., Das, R.K., Qian, X., Shou, M.Z., Li, H.: Is someone speaking?: Exploring long-term temporal features for audio-visual active speaker detection. In: Proceedings of the 29th ACM International Conference on Multimedia (ACM MM). p. 3927–3935 (2021)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
50. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing **37**(3), 328–339 (1989)
51. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14549–14560 (2023)

52. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6023–6032 (2019)
53. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (ICLR) (2018)
54. Zhang, W., Wang, J., Luo, Y., Yu, L., Yu, W., He, Z., Shen, J.: Mtga: Multi-view temporal granularity aligned aggregation for event-based lip-reading. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2025)
55. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–21. Springer (2022)
56. Zhu, Q., Zhou, L., Zhang, Z., Liu, S., Jiao, B., Zhang, J., Dai, L., Jiang, D., Li, J., Wei, F.: Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia* **26**, 1–13 (2023)