Enroll Once, Identify Everywhere: Few-Shot Cross-Scenario Visual Speaker Recognition via Event-driven SlowFast Networks

Junguang Yao[1], Peichun Hua[2], and Yue Zheng[1]✉

[1] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, Guangdong, 518172, P. R. China
junguangyao@link.cuhk.edu.cn, zhengyue@cuhk.edu.cn

[2] School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, Guangdong, 518172, P. R. China
peichunhua@link.cuhk.edu.cn

Abstract. Visual Speaker Recognition (VSR) based on lip motion provides a silent biometric that is robust to ambient acoustic noise, but practical deployment requires recognizing newly enrolled users from only a few samples under conditions that may differ from enrollment. We study this problem with event cameras under a deployment-oriented few-shot enrollment protocol, where training and test identities are disjoint, each unseen test identity contributes three enrollment samples, and gallery and probe scenarios may differ in viewpoint or illumination. We adapt a SlowFast backbone to two-channel polarity-aware event count volumes, using complementary temporal pathways and fast-to-slow lateral fusion to model lip articulation. The network is trained with cross-entropy on training identities and used as a frozen feature extractor for nearest-neighbor identification at test time. On DVSpeaker, the model achieves 99.18% Rank-1 in matched frontal-bright conditions and 87.25%, 74.95%, and 50.20% under low-light, 45°, and 90° shifts, respectively. Ablations show that the two temporal pathways are complementary, while lateral fusion provides its clearest gains under viewpoint shifts. These results demonstrate a practical event-based formulation for few-shot cross-scenario VSR without retraining for newly enrolled identities. Our source code is available on https://github.com/JiuZeongit/DVSpeaker4Identification.

Keywords: Event camera · Visual speaker recognition · Lip motion · SlowFast network · Few-shot enrollment · Cross-scenario biometrics.

1 Introduction

Visual Speaker Recognition (VSR) seeks to recognize an individual from the visible dynamics of speech articulation, most prominently the motion of the lips and perioral region. Compared with acoustic speaker recognition, VSR is silent, contactless, and naturally robust to ambient acoustic noise, which makes it appealing for unlocking mobile devices, controlling access in shared environments,

and providing a secondary cue in multimodal authentication [2,8,9]. Among the visual cues that can be extracted from a speaking face, lip motion in particular has been shown to encode behavioral characteristics that are difficult to imitate or synthetically forge, and that complement static facial appearance [8,17].

Despite this potential, conventional frame-based cameras struggle to capture the fine-grained temporal structure of articulation. Their fixed integration windows blur fast lip openings and closings, and their limited dynamic range degrades quickly under low light. Event cameras circumvent both limitations by reporting per-pixel brightness changes asynchronously with microsecond latency and across a wide dynamic range [6]. Recent works [12,14,21] have shown that the asynchronous polarity stream of an event sensor preserves the spatiotemporal patterns required for speaker recognition while remaining largely insensitive to ambient illumination, which is precisely the regime in which appearance-based methods are most fragile.

Two further requirements, however, must be met before lip-motion VSR is usable in practice. First, an end user expects to enroll once, with at most a handful of utterances, and to be recognized reliably afterward: there is no opportunity to retrain the model for every new identity. Second, enrollment is typically performed under benign conditions (frontal pose, good lighting), but recognition must succeed when the deployment context drifts to off-axis viewpoints or low-light environments. These two requirements jointly define what we will refer to as the *few-shot cross-scenario* setting. Most prior lip-motion studies handle each requirement in isolation: matched-scenario benchmarks assume identical training and evaluation conditions, while strict cross-scenario protocols, such as NeuroLip [21], train and evaluate on a closed identity set across different scenarios. Although such protocols measure robustness to scenario shift, they do not capture deployment-time enrollment, where the recognition model is first deployed and then new users, unseen during training, are enrolled with only a few samples. Consequently, neither protocol fully reflects the operating condition of a real-world biometric system.

In this paper, we study a complementary, deployment-oriented cross-scenario identification protocol. The protocol conceptually comprises three distinct phases: training, enrollment, and probing. During training, network parameters are optimized on a pool of training identities across diverse environmental conditions (e.g., varying viewpoints and illumination), which enables the network to discover scenario-invariant identity cues. At test time, the trained network acts purely as a feature extractor for a disjoint pool of unseen users. During enrollment, a few frictionless samples are collected from the registered users to form an enrollment gallery. These registered users can be identified against the gallery during probing via simple nearest-neighbor search on ℓ_2-normalized embeddings. Crucially, this setup ensures that while the model learns robust representations from diverse data, the actual deployment (enrollment and probing) can seamlessly occur in arbitrary, possibly mismatched conditions.

To match this setting on the architectural side, we adopt a SlowFast [5] backbone tailored to events. The intuition is that lip motion contains both slowly

varying postural information, such as lip shape and tongue positioning, as well as fast transients, such as rapid lip closure and reopening. An event volume already provides excellent temporal resolution for the latter. We instantiate two 3D residual pathways that share the same input but operate at different temporal rates, and we connect them by a fast-to-slow lateral pathway at every fusion stage so that the slow stream is continuously enriched with high-rate motion features. The final embedding concatenates global descriptors from both pathways.

Our contributions are summarized as follows. First, we formalize the few-shot cross-scenario VSR setting, which combines disjoint train-test identities with an enrollment of only three samples per probe identity, and we instantiate it on the DVSpeaker [21] event-camera corpus across four perspectives and illumination scenarios. Second, we present an event-driven SlowFast backbone that takes a two-channel polarity-aware count volume and is trained end-to-end with cross-entropy, with no auxiliary metric-learning loss. Third, we report a thorough empirical study of how training-pool size, enrollment scenario, and probing scenario interact, including a 4×4 cross-condition matrix and a component-wise ablation that isolates the contribution of each pathway and the fusion mechanism. Under the most favorable comparable setting (30 training and 20 testing identities, with frontal-bright enrollment), our method reaches 99.18% Rank-1 in matched conditions, 87.25% under low-light frontal probing, 74.95% under 45° probing, and 50.20% under 90° probing, substantially exceeding state-of-the-art event-based baselines reproduced under the identical protocol.

2 Related Work

2.1 Lip-based VSR

Early lip biometric studies focused on physiological cues such as lip texture and shape, modeled with handcrafted descriptors [15,17]. With the advent of deep models, attention shifted to behavioral cues: the spatiotemporal patterns of articulation during speech. Frame-based pipelines such as LipAuth [9] and DynamicLip [2] exploit dynamic features extracted from RGB video, often jointly with appearance, and report near-perfect accuracy when the training and test conditions match. He et al. [8] explicitly disentangle static identity, dynamic identity, and content, demonstrating that identity-specific motion is more robust to scene variation than appearance. Alternative sensors, including acoustic systems such as LipPass [11] and radar-based pipelines such as Lip-TWUID [20], have also been explored for lip-based authentication. They sidestep some illumination issues but require specialized hardware and are sensitive to environmental noise or motion artifacts.

To overcome the limitation of conventional sensors, event cameras have begun to receive attention in biometric applications. Early studies [13,18] explored event cameras for biometric identification using gait cues. Moreira et al. [12] further presented the first event-based method for identity recognition from facial dynamics; NeuroLip [21] then introduced the DVSpeaker event corpus and

proposed a learnable temporal voxel encoding combined with polarity-aware regularization, evaluated under a closed-set cross-scenario protocol.

None of the existing event-based lip-motion VSR studies, however, directly address deployment-time enrollment, where enrolled users are unseen during training and only a few enrollment samples are available. Most also evaluate under matched enrollment and probing conditions, rather than testing robustness to viewpoint or illumination shifts. To bridge these gaps, the present work utilizes the DVSpeaker dataset but fundamentally departs from NeuroLip's closed-identity paradigm. Specifically, we target the identification of *unseen* identities with a strict few-shot enrollment and adopt a generic event-driven SlowFast backbone whose two pathways are designed to model multi-rate articulation dynamics.

2.2 SlowFast Architectures and Event-based Classification

SlowFast networks [5] were originally introduced for action recognition in RGB video. They process the same input through two pathways that operate at different temporal rates: a low-rate Slow pathway that captures spatially detailed but temporally coarse semantics and a high-rate Fast pathway that emphasizes motion. Lateral connections from Fast to Slow allow the heavier Slow stream to integrate motion cues without inflating its width. The design is appealing for event data because event volumes are intrinsically high-rate and motion-dominated, while lip articulation is multi-scale in time. WhisperNetV2 [22] previously combined a Siamese SlowFast with RGB lip video for biometrics; we instead apply a single-branch SlowFast directly to event count volumes and rely on cross-entropy training with disjoint identity splits.

Event-based classification methods can be grouped by how they convert the asynchronous stream into a tensor consumable by deep models. Spiking networks process raw events directly [4] but suffer from training instability. Voxel and time-surface representations [1,7,10] convert events into dense tensors compatible with standard CNNs, but this conversion discretizes the event stream into fixed temporal bins or surfaces and may therefore discard part of the sensor's native temporal precision. Graph-based and point-cloud-based encoders [3,16] preserve sparsity but are sensitive to spurious events. Instead, we adopt a simple two-channel polarity-aware count volume with optional logarithmic compression, motivated by its robustness, low pre-processing cost, and direct compatibility with 3D convolutions.

2.3 Cross-identity and Few-shot Recognition

Identity recognition with disjoint training and testing identities is standard in face and gait recognition, where enrolled users are typically not part of the training pool. The accepted recipe is to train a classifier or a metric on a closed pool, discard the classification head at deployment, and use the penultimate representation as a generic embedding for nearest-neighbor matching against a small gallery. This protocol underpins both classical face recognition pipelines and

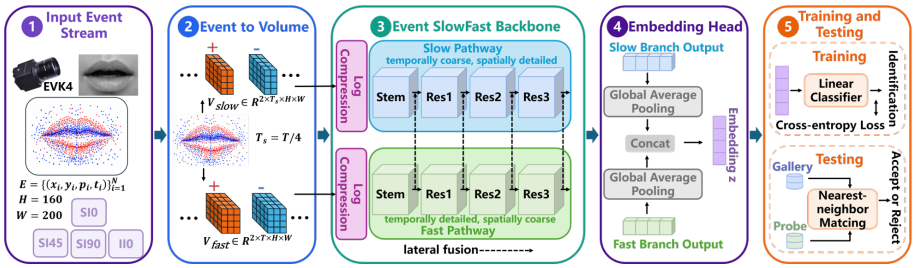

Fig. 1. Overall framework of the proposed event-based lip-motion identification method. The asynchronous event stream is converted into polarity-aware slow and fast count volumes, processed by an Event SlowFast backbone, and mapped to an identity embedding for training and gallery–probe identification.

contemporary deep approaches [19]. Our setting follows the same logic but adds two specific constraints: the gallery contains only three samples per identity, and the gallery and probe scenarios may differ. Although few-shot and cross-identity lip-based biometric authentication have been studied with conventional video, their combination with event-based lip-motion sensing and cross-scenario gallery–probe matching remains largely unexplored. To the best of our knowledge, our work is the first to systematically evaluate disjoint-identity few-shot enrollment on an event-based lip-motion corpus under explicit viewpoint and illumination shifts.

3 Method

Fig. 1 presents the overall framework of the proposed method. Given an asynchronous lip-motion event stream, we first convert the events into polarity-aware count volumes with log-count compression, forming separate slow and fast inputs for the subsequent backbone. The Event SlowFast backbone then extracts complementary spatiotemporal representations: the Slow pathway focuses on spatially detailed but temporally coarse patterns, while the Fast pathway captures fine-grained temporal dynamics and provides lateral motion cues to the Slow pathway. The outputs of the two pathways are aggregated by global average pooling and concatenated into an identity embedding. During training, the embedding is supervised by a linear classifier with cross-entropy loss; during evaluation, the classifier is discarded, and ℓ_2-normalized embeddings are used for nearest-neighbor gallery–probe identification without test-time finetuning.

3.1 Problem Formulation

This work explores a few-shot enrollment setting, where K-shot denotes that each unseen test identity contributes only K enrollment samples to construct the gallery. The encoder itself is trained conventionally on the training identity

pool and is neither meta-trained nor fine-tuned on the test identities. Specifically, let $\mathcal{D}$ denote an event-based corpus, where each sample x is associated with an identity label $y \in \mathcal{Y}$ and a scenario label $s \in \mathcal{S}$. The scenarios cover various viewpoints and illumination conditions. We split $\mathcal{Y}$ into a training pool $\mathcal{Y}_{\text{tr}}$ and a disjoint testing pool $\mathcal{Y}_{\text{te}}$, with $\mathcal{Y}_{\text{tr}} \cap \mathcal{Y}_{\text{te}} = \emptyset$. A backbone f_θ is trained on samples drawn from $\mathcal{Y}_{\text{tr}}$ over all scenarios. At evaluation time, given a gallery scenario $s_g \in \mathcal{S}$ and a probe scenario $s_p \in \mathcal{S}$, the gallery is constructed by sampling exactly $K = 3$ samples per test identity from s_g. The probe set then comprises the remaining samples from s_p if $s_p = s_g$ (matched-scenario) or all available samples of these identities from s_p if $s_p \neq s_g$ (cross-scenario). A probe is correctly identified at Rank-1 if its nearest gallery neighbor, measured by Euclidean distance on ℓ_2-normalized embeddings, shares the same identity. The reported metric throughout the paper is Rank-1 accuracy on the probe set.

This deployment-time few-shot enrollment setting is strict in three respects. The training and testing identities are disjoint, so the network never sees the test users. The gallery is small (K=3), which simulates a frictionless enrollment phase. And s_g and s_p may be arbitrarily chosen, so the same backbone must support both matched-scenario and cross-scenario recognition without any test-time finetuning.

3.2 Event-to-Volume Representation

Given an event segment $E = \{(x_i, y_i, p_i, t_i)\}_{i=1}^{N}$ with spatial coordinates $(x_i, y_i) \in [0, W) \times [0, H)$, polarity $p_i \in \{0, 1\}$, and timestamp t_i, we construct a four-dimensional count volume $V \in \mathbb{R}^{2\times T\times H\times W}$ by linearly mapping the timestamp range of the segment onto T uniform temporal bins and accumulating event counts per polarity. To stabilize the dynamic range across illumination conditions, we apply an element-wise logarithmic compression, $V \leftarrow \log(1 + V)$. The two polarity channels are kept separate throughout the network because the spatial distribution of positive and negative events jointly reflects local motion direction.

3.3 Event SlowFast Backbone

We adopt a SlowFast [5] design with two parallel 3D residual pathways. The Slow pathway operates on a temporally subsampled volume $V_{\text{slow}} \in \mathbb{R}^{2\times T_s\times H\times W}$ with $T_s = T/4$, while the Fast pathway operates on $V_{\text{fast}} \in \mathbb{R}^{2\times T\times H\times W}$. Both pathways share the same architectural skeleton: a stem followed by three residual stages, with progressive spatial downsampling but a clear separation of temporal behavior. Here, the stem denotes the initial feature extraction module in each pathway, consisting of two 3D convolutional blocks. The first block reduces spatial resolution with a $3 \times 5 \times 5$ convolution of stride $(1, 2, 2)$, while the second $3 \times 3 \times 3$ block further refines early spatiotemporal features. Each residual stage contains two 3D residual blocks, where the input is added back to the convolutional output through either an identity shortcut or a $1 \times 1 \times 1$ projection

when channel or resolution changes. This design supports deeper feature extraction while preserving stable information flow. To keep compute manageable, we adopt the asymmetric channel capacity design inspired by the original SlowFast principle: $C_0^{\text{slow}} = 32$ versus $C_0^{\text{fast}} = 16$ at the stem, doubling at every subsequent stage. Equipped with the wider channel base and by preserving T_s throughout (the temporal stride at every stage is one), the Slow pathway yields features that are spatially detailed but temporally summarized. In contrast, the Fast pathway relies on its narrower channel base to process the full frame rate efficiently. It halves the temporal extent at the second and third stages, producing features that are temporally fine-grained but spatially compact.

Lateral fusion. At four points in the network (after the stem and after each residual stage), Fast features are projected to the Slow channel dimension by a $1 \times 1 \times 1$ convolution and are temporally aligned to T_s via adaptive average pooling along the temporal axis. The aligned features are then merged into the Slow stream. We default to additive fusion, since it requires no extra parameters once the projection is done, and we report a concatenation variant in the ablation. Formally, let $k \in \{0, 1, 2, 3\}$ index the four fusion points. With ϕ denoting projection, Π_T adaptive temporal pooling to T_s, and σ a ReLU activation, the fusion is defined as:

$$V_{\text{slow}}^{(k+1)} = \sigma\Big(V_{\text{slow}}^{(k)} + \phi\Big(\Pi_T\Big(V_{\text{fast}}^{(k)}\Big)\Big)\Big), \tag{1}$$

which makes the Slow pathway a temporally-coarse but motion-informed branch by construction. There is no Slow-to-Fast lateral, mirroring the original design.

Embedding head. At the end of both pathways we apply 3D global average pooling and concatenate the resulting vectors to form the embedding $z \in \mathbb{R}^{384}$ (256 from Slow, 128 from Fast). During training, z is fed to a single linear classifier of size $|\mathcal{Y}_{\text{tr}}|$ supervised by cross-entropy. At test time, the classifier is discarded; z is ℓ_2-normalized and used directly for nearest-neighbor identification.

3.4 Training and Testing

The proposed network acts purely as a feature extractor. We deliberately avoid auxiliary metric losses during training; instead, cross-identity generalization is achieved simply by training a sufficiently expressive multi-condition classifier on disjoint identities and subsequently reusing its embeddings. We attribute this to the SlowFast representation already disentangling identity-discriminative motion from scenario-specific appearance, an effect we measure quantitatively in the ablation.

Enrollment of a new test identity consists of (i) computing the embedding z for each of its three gallery samples, (ii) applying ℓ_2 normalization, and (iii) storing them in the gallery. There is no per-identity finetuning, no prototype averaging, and no calibration step. A probe is matched to its nearest gallery embedding under Euclidean distance, which is monotonic with cosine similarity for normalized vectors.

4 Experiments

4.1 Dataset, Splits, and Implementation

We use DVSpeaker [21], an event-camera lip-motion dataset of 50 subjects articulating digits 0–9, recorded at 200×160 pixels under four scenarios (SI0, SI45, SI90, II0), including three viewpoints (0°, 45°, 90°) under sufficient illumination (SI, ∼216 lux), and a frontal view (0°) under insufficient illumination (II, ∼12.5 lux). Each subject contributes 100 samples per scenario, for a total of 20,000 segments. Following the few-shot cross-scene protocol of Sec. 3, we split the 50 identities into disjoint training and testing pools. We consider three pool sizes (10/40, 20/30, 30/20 train/test) to study how the size of the training pool influences open-set generalization; the 30/20 split is our default and is used for all subsequent ablations and comparisons.

Training samples cover all four scenarios; the test pool is reused for every (gallery, probe) scenario pair. All experiments use $K = 3$ enrollment samples per test identity, drawn deterministically by a per-identity seed so that the gallery and probe samples are mutually exclusive. We implement the model in PyTorch and train on a single NVIDIA GeForce RTX 4090 GPU. We use Adam with a learning rate of 1×10^{-3}, a batch size of 16, and 50 epochs, and we monitor validation Rank-1 averaged over the four matched-scene evaluations on a held-out portion of the training identities. The best checkpoint is selected based on this average and is then used for all subsequent evaluations. Additionally, we use $W = 200$, $H = 160$, and $T = 32$ in all experiments unless otherwise noted; the choice follows the spatial crop introduced when DVSpeaker was curated [21], and is small enough to allow fast 3D processing while preserving lip detail.

4.2 Matched-Scenario Identification

Table 1 reports Rank-1 accuracy in the matched-scenario setting, where the gallery and probe scenarios are identical, i.e., $s_g = s_p$ for different sizes of the training identity pool. With only ten training identities, the network already exceeds 74% in every scenario, indicating that the SlowFast embedding generalizes non-trivially to unseen users with very little training material. Performance scales smoothly with the size of the training pool: at 30 identities, Rank-1 is at or above 96% in every scenario. Interestingly, the highest matched-scenario score is consistently observed on SI90 when the training pool is small (10 or 20 identities), even though SI90 is the most extreme viewpoint. We attribute this to the strong inter-subject variability of the lateral profile: when both gallery and probe come from SI90, the same idiosyncratic lip-and-jaw silhouette is observed twice, which is highly discriminative even with three enrollment samples. As the training pool grows, this advantage flattens out, and the four scenarios converge.

4.3 Cross-Scenario Identification

Table 2 reports the full 4×4 matrix of (gallery, probe) scenario combinations under the default 30/20 split. The diagonal corresponds to matched scenarios (and

Table 1. Matched-scenario Rank-1 accuracy (%) on DVSpeaker as a function of the training identity pool size. The remaining identities form the test pool. K=3 enrollment samples per test identity. Diagonal entries: s_g=s_p.

Train / Test identities	SI0	SI45	SI90	II0
10 / 40	78.22	76.13	85.41	74.82
20 / 30	91.68	92.44	94.50	89.90
30 / 20	**99.18**	**97.32**	**98.04**	**96.34**

reproduces the bottom row of Table 1); off-diagonal entries quantify the cost of a scenario shift between enrollment and probing. Three observations stand out. First, illumination shifts at the same viewpoint (SI0↔II0) are the most benign, with Rank-1 staying around 87–91%. Even though the absolute event count is dramatically lower under II0, the polarity-aware count volume preserves the spatial structure of articulation, and the SlowFast model learns embeddings that depend on this structure rather than on absolute event density. Second, the 45° shift (SI0↔SI45) costs roughly 22–25 percentage points relative to the matched diagonal. Third, the 90° shift (SI0↔SI90) is the hardest, halving accuracy to about 50%. The cross-condition matrix is approximately, but not exactly, symmetric. This is expected because gallery and probe scenarios play different roles under few-shot nearest-neighbor identification: only three samples per identity define the gallery, while the probe scenario determines the distribution of test samples to be classified. Consequently, reversing a scenario pair can change Rank-1 accuracy, although the direction of the change is scenario-dependent rather than uniformly attributable to gallery noise or probe noise.

Table 2. Cross-scenario Rank-1 accuracy (%) on DVSpeaker, 30 training and 20 testing identities. Each row is a gallery scenario, each column a probe scenario. Diagonal: matched-scenario.

Gallery ↓ / Probe →	Probe scenario			
	SI0	SI45	SI90	II0
SI0	**99.18**	74.95	50.20	87.25
SI45	77.45	**97.32**	58.75	70.05
SI90	50.80	53.10	**98.04**	35.75
II0	90.90	65.90	39.00	**96.34**

A similar pattern, with consistently lower absolute numbers, is observed under the more demanding 20/30 split (Table 3). The relative drop from the matched diagonal to the 90° cross-shift is comparable to the 30/20 case (about 50 percentage points), suggesting that the failure mode of orthogonal-view enrollment is structural rather than data-bound: the lateral silhouette simply does

not preserve enough articulation cues observable from the frontal view, regardless of how many training identities we have.

Table 3. Cross-scenario Rank-1 accuracy (%) on DVSpeaker, 20 training and 30 testing identities.

Gallery ↓ / Probe →	Probe scenario			
	SI0	SI45	SI90	II0
SI0	**91.68**	55.33	35.47	67.77
SI45	54.10	**92.44**	40.23	41.77
SI90	35.20	45.63	**94.50**	25.87
II0	78.50	51.67	30.93	**89.90**

4.4 Comparison with Existing Methods

We retrain four representative event-based identity-recognition baselines on the same 30/20 split and evaluate them under the same gallery-probe protocol with $K = 3$. We use the official source code where available and adapt it minimally to consume the DVSpeaker preprocessing pipeline. The first baseline, proposed by Moreira et al. [12] (denoted as EventFace in our comparison), is the original full-face event-based identification work and represents short-term facial motion modeling. The second, EV-Gait-IMG [18], is a strong event-stream gait baseline that uses image-like accumulated frames with a deep CNN. The third, which is denoted as ANN in [4], is a custom frame-rate convolutional baseline trained on the same count volumes. The fourth, SpikGRU++ [4], is a spiking-network method originally proposed for event-based lip reading and is the only baseline in our pool that consumes raw events without densification. For a fair and practical comparison, we use the SI0 gallery row, which is the deployment-friendly enrollment scenario, as the basis for comparison. Results are summarized in Table 4.

The matched-scenario number (SI0→SI0) is saturated for the densified representation baselines (EventFace, EV-Gait-IMG), but it is markedly lower for SpikGRU++ (67.42%), reflecting a well-documented training-stability gap of spiking networks on event-based identity tasks [4]. The picture changes sharply on cross-scenario probes. Our SlowFast model leads by 28.75 points on SI45, 21.15 points on SI90, and 8.55 points on II0 over the strongest baseline at each scenario. The gap is widest at viewpoint shifts, where modeling motion at multiple temporal scales appears to substantially help, and is narrowest at the illumination shift, where the polarity-aware count volume is already a strong inductive bias.

4.5 Ablation Studies

We isolate the contribution of each pathway and the lateral fusion by training three reduced variants under the 30/20 split with frontal-bright enrollment. *Slow*

Table 4. Comparison with reproduced event-based baselines under our few-shot cross-scenario protocol (30 train / 20 test, gallery=SI0, K=3). Values are Rank-1 accuracy (%) on the indicated probe scenario.

Method	SI0	SI45	SI90	II0
EventFace [12]	97.57	41.00	29.05	78.70
EV-Gait-IMG [18]	92.32	26.50	21.75	60.40
ANN [4]	77.16	46.20	16.10	57.70
SpikGRU++ [4]	67.42	43.05	20.85	47.80
Event SlowFast (ours)	**99.18**	**74.95**	**50.20**	**87.25**

only keeps the Slow pathway and routes its embedding directly to the classifier; *Fast only* keeps the Fast pathway under the same recipe; and *SlowFast w/o fusion* retains both pathways but disables the fast-to-slow lateral connections, so that the two streams are trained as parallel encoders and are concatenated only at the embedding stage. Results are summarized in Table 5.

Table 5. Ablation of the SlowFast components under 30 train / 20 test, gallery=SI0, K=3. Rank-1 accuracy (%) per probe scenario.

Variant	SI0	SI45	SI90	II0
Fast only	97.88	59.95	28.00	87.45
Slow only	98.29	58.00	31.10	86.15
SlowFast w/o fusion	98.45	64.60	41.90	**89.85**
SlowFast (full)	**99.18**	**74.95**	**50.20**	87.25

First, neither pathway is sufficient on its own. Slow only and Fast only variants yield very similar matched-scenario performance (∼98%) but lag the full model by 15–17 points on the 45° and 22–19 points on the 90° cross-scenario, indicating that the strong cross-scenario gain comes from the combination, not from either branch in isolation. Second, the parallel-without-fusion variant already recovers some of the gap (about 4–11 points over the better single pathway on SI45 and SI90), suggesting that the slow and fast features are genuinely complementary at the embedding level. Third, the lateral fusion adds another 8–10 points on the viewpoint-shifted scenarios, indicating that propagating fast-stream motion cues into the slow stream is particularly beneficial when enrollment and probing viewpoints differ. In contrast, under the illumination-only shift (SI0→II0), the full model obtains 87.25%, compared with 89.85% without lateral fusion. This suggests that the benefit of lateral fusion is scenario-dependent and is primarily associated with viewpoint robustness rather than illumination robustness. The picture is essentially preserved at the 20/30 split, although absolute numbers are lower; for brevity, we omit the full table and note that the relative ranking of the four variants is the same.

We also tested concatenation as an alternative to additive fusion, with the post-projection 1×1×1 convolution adjusted accordingly. The two variants are within 0.6 points on every probe scenario; given the negligible gain and the additional parameters required, we keep the additive fusion as the default.

4.6 Discussion

Where the model wins, and where it does not. The strongest cross-scenario gains, both in absolute terms and relative to baselines, are observed at the 45° viewpoint shift and at the low-light shift. Both are scenarios where the absolute appearance of the lip region differs from the gallery, but the underlying articulation *trajectory* is preserved, and a multi-rate motion model is in a position to exploit that trajectory. The 90° shift is qualitatively different: the lateral profile occludes the inner mouth and removes the temporal cues that the network learned to attend to under frontal-like training. We observe in Tables 2–3 that even when 90° is the gallery scenario, accuracy on SI0 and SI45 probes is also depressed, which is consistent with a structural rather than statistical limitation.

Comparison with prior NeuroLip protocol. It is tempting to read across to the cross-scenario numbers reported in NeuroLip [21], where the strict SI0-only training setting collapsed on SI90 but performed well on II0 and SI45. The two protocols are not directly comparable: NeuroLip is a closed-set identification problem that keeps the 50 identities fixed and varies only the *scenarios*, whereas this work is a cross-identity problem that exposes the encoder to all four scenarios during training and varies the *identities*. For practitioners who can afford to label data across scenarios but want to avoid constant retraining of the model when new users enroll, our protocol is the more practical operating point, and the SlowFast architecture effectively exploits multi-condition training.

Limitations. We do not study cross-dataset generalization in this paper, since DVSpeaker is, to our knowledge, the only event-camera lip-motion corpus with controlled multi-scenario coverage; cross-dataset evaluation against the public DVSLip [14] (matched-scenario and word-based) would mainly probe content invariance rather than scenario robustness and is left for future work. Finally, the corpus contains 50 healthy adult speakers; a larger and more demographically diverse pool would be needed for a deployment-grade biometric study.

5 Conclusion

We present a few-shot cross-scenario formulation of VSR on event cameras, in which a SlowFast 3D backbone is trained on disjoint identities under all available scenarios and is then frozen and used purely as a feature extractor for nearest-neighbor identification with three enrollment samples per new user. On the DVSpeaker [21] dataset, this combination reaches 99.18% Rank-1 on matched frontal-bright probing, 87.25% under low-light frontal probing, and 74.95% under a 45°

viewpoint shift, substantially exceeding four event-based baselines. Ablations demonstrate the complementarity of the dual-rate pathways, while fast-to-slow lateral fusion provides its clearest additional gains under viewpoint shifts. The remaining failure mode is the orthogonal 90° shift, which we attribute to structural occlusion of the articulation trajectory and not to data scarcity. Future work includes scenario-conditional fusion strategies, geometry-aware backbones for extreme viewpoints, and a study of whether the same SlowFast embedding transfers to non-digit articulation content.

Acknowledgement This research is supported by the Shenzhen Stability Science Program 2023.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baldwin, R.W., Liu, R., Almatrafi, M., Asari, V., Hirakawa, K.: Time-ordered recent event (TORE) volumes for event cameras. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2519–2532 (2023). https://doi.org/10.1109/TPAMI.2022.3172212
2. Chen, H., Xu, Y., Feng, Y., Jian, M., Liu, F., Hu, P., Peng, K., He, S., Wang, Z.: DynamicLip: Shape-independent continuous authentication via lip articulator dynamics. IEEE Trans. Mob. Comput. **25**(3), 4063–4076 (2026). https://doi.org/10.1109/TMC.2025.3624155
3. Chen, L., Zhang, Z., Xiao, Y., Wang, Y.: EGST: An efficient solution for human gaits recognition using neuromorphic vision sensor. IEEE Trans. Inf. Forensics Security **19**, 6144–6154 (2024). https://doi.org/10.1109/TIFS.2024.3409167
4. Dampfhoffer, M., Mesquida, T.: Neuromorphic lip-reading with signed spiking gated recurrent units. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 2141–2151 (2024). https://doi.org/10.1109/CVPRW63382.2024.00219
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 6201–6210 (2019). https://doi.org/10.1109/ICCV.2019.00630
6. 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 Trans. Pattern Anal. Mach. Intell. **44**(1), 154–180 (2022). https://doi.org/10.1109/TPAMI.2020.3008413
7. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 5632–5642 (2019). https://doi.org/10.1109/ICCV.2019.00573
8. He, Y., Yang, L., Wang, S., Liew, A.W.C.: Lip feature disentanglement for visual speaker authentication in natural scenes. IEEE Trans. Circuits Syst. Video Technol. **34**(10), 9898–9909 (2024). https://doi.org/10.1109/TCSVT.2024.3405640
9. Kuang, L., Zeng, F., Liu, D., Cao, H., Jiang, H., Liu, J.: LipAuth: Securing smartphone user authentication with lip motion patterns. IEEE Internet Things J. **11**(1), 1096–1109 (2024). https://doi.org/10.1109/JIOT.2023.3289573

10. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.: HOTS: A hierarchy of event-based time-surfaces for pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(7), 1346–1359 (2017). https://doi.org/10.1109/TPAMI.2016.2574707
11. Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Liu, Y., Li, M.: LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In: Proc. IEEE Conf. Comput. Commun. (INFOCOM). pp. 1466–1474 (2018). https://doi.org/10.1109/INFOCOM.2018.8486283
12. Moreira, G., Graça, A., Silva, B., Martins, P., Batista, J.P.: Neuromorphic event-based face identity recognition. In: Proc. Int. Conf. Pattern Recognit. (ICPR). pp. 922–929 (2022). https://doi.org/10.1109/ICPR56361.2022.9956236
13. Sokolova, A., Konushin, A.: Human identification by gait from event-based camera. In: Proc. Int. Conf. Mach. Vis. Appl. (MVA). pp. 1–6 (2019). https://doi.org/10.23919/MVA.2019.8758019
14. 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: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 20062–20071 (2022). https://doi.org/10.1109/CVPR52688.2022.01946
15. Tsuchihashi, Y.: Studies on personal identification by means of lip prints. Forensic Sci. **3**(3), 233–248 (1974). https://doi.org/10.1016/0300-9432(74)90034-X
16. Wang, Q., Zhang, Y., Yuan, J., Lu, Y.: Space-time event clouds for gesture recognition: From RGB cameras to event cameras. In: Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV). pp. 1826–1835 (2019). https://doi.org/10.1109/WACV.2019.00199
17. Wang, S.L., Liew, A.W.C.: Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power. Pattern Recognit. **45**(9), 3328–3335 (2012). https://doi.org/10.1016/J.PATCOG.2012.02.016
18. Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Cui, L., Wen, H.: Event-stream representation for human gaits identification using deep neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3436–3449 (2022). https://doi.org/10.1109/TPAMI.2021.3054886
19. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Trans. Pattern Anal. Mach. Intell. **39**(2), 209–226 (2017). https://doi.org/10.1109/TPAMI.2016.2545669
20. Yang, K., Zhu, D., Han, C., Guo, J., Sun, S., Sun, L.: Lip-TWUID: Noninvasive through-wall user identification using SISO radar and lip movement micro-doppler signatures with limited samples. IEEE Trans. Instrum. Meas. **74**, 1–17 (2025). https://doi.org/10.1109/TIM.2025.3541807
21. Yao, J., Liu, W., Picek, S., Zheng, Y.: NeuroLip: An event-driven spatiotemporal learning framework for cross-scene lip-motion-based visual speaker recognition (2026). https://doi.org/10.48550/ARXIV.2604.15718
22. Zakeri, A., Hassanpour, H., Khosravi, M.H., Nourollah, A.M.: WhisperNetV2: SlowFast siamese network for lip-based biometrics. arXiv preprint arXiv:2407.08717 (2024). https://doi.org/10.48550/ARXIV.2407.08717