Finger in Camera Speaks Everything: Unconstrained Air-Writing for Real-World

Meiqi Wu *Member, IEEE*, Kaiqi Huang *Senior Member, IEEE*, Yuanqiang Cai, Shiyu Hu, Yuzhong Zhao *Member, IEEE*, Weiqiang Wang* *Member, IEEE*

***Abstract*—Air-writing is a challenging task that combines the fields of computer vision and natural language processing, offering an intuitive and natural approach for human-computer interaction. However, current air-writing solutions face two primary challenges: (1) their dependency on complex sensors ($e.g.$, Radar, EEGs and others) for capturing precise handwritten trajectories, and (2) the absence of a video-based air-writing dataset that covers a comprehensive vocabulary range. These limitations impede their practicality in various real-world scenarios, including the use on devices like iPhones and laptops. To tackle these challenges, we present the groundbreaking air-writing Chinese character video dataset (AWCV-100K), serving as a pioneering benchmark for video-based air-writing. This dataset captures handwritten trajectories in various real-world scenarios using commonly accessible RGB cameras, eliminating the need for complex sensors. AWCV-100K includes 8.8 million video frames, encompassing the complete set of 3,755 characters from the GB2312-80 level-1 set (GB1). Furthermore, we introduce our baseline approach, the video-based character recognizer (VCRec). VCRec adeptly extracts fingertip features from sparse visual cues and employs a spatio-temporal sequence module for analysis. Experimental results showcase the superior performance of VCRec compared to existing models in recognizing air-written characters, both quantitatively and qualitatively. This breakthrough paves the way for enhanced human-computer interaction in real-world contexts. Moreover, our approach leverages affordable RGB cameras, enabling its applicability in a diverse range of scenarios. The code and data examples will be made public at https://github.com/wmeiqi/AWCV.**

***Index Terms*—Air-writing, real-world, benchmark, video-based air-writing Chinese character recognition.**

I. Introduction

WHEN you gesture with your fingertips to write characters in front of the camera, advanced artificial intelligence technology instantly grasps your intentions and initiates the command execution. This concept, commonly referred to as *"finger in camera speaks everything"*, epitomizes the essence of air-writing as a powerful technique facilitating effective human-computer interaction. It entails the recognition of characters from handwritten trajectories in 3D space, resembling depictions seen in movies like Ready Player One.

As depicted in Fig. 1 (middle), air-writing holds broad application prospects, such as its integration into popular smart devices ($e.g.$, VR/AR/MR devices, Smart TV, Laptop, iPhone, iPad), enabling machines to interpret human intentions in a contactless and silent manner. Specifically, air-writing can be used in intelligent conversational systems ($e.g.$, GPT series [6], ChatGLM [7]) or the metaverse. Given the diverse nature

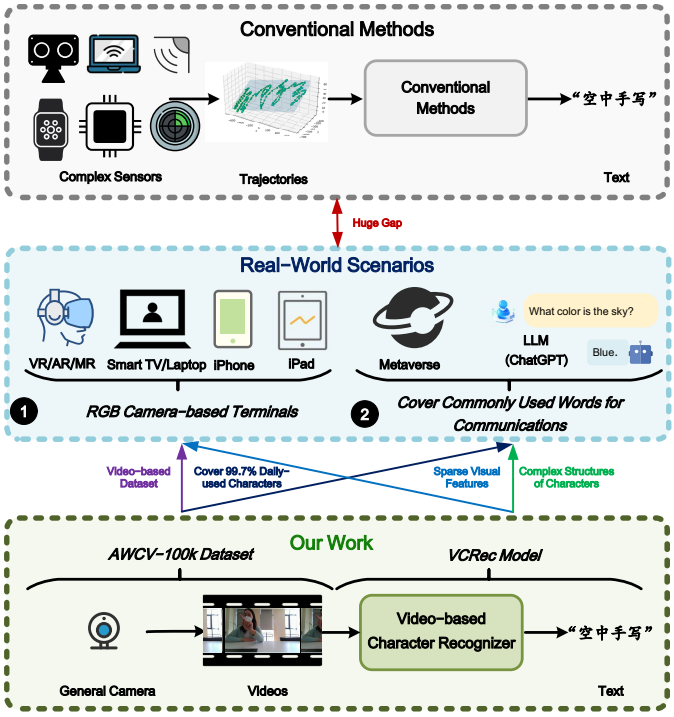


Fig. 1. Comparing Our Work and Conventional Air-Writing in Real-World Scenarios. Conventional air-writing relies on accurately captured handwritten trajectories by complex sensors (such as Radar [1], Smart Watch [2], Leap Motion [3], EEGs [4], IMU [5]), which impose significant limitations for real-world scenarios ($e.g.$, VR/AR/MR, iPhone, metaverse, GPT series [6] and others). Mainstream real-world devices only incorporate standard RGB cameras and require coverage of commonly used words for communication purposes. To address these challenges, we propose a video-based air-writing dataset with a comprehensive corpus (covering 99.7% of daily-used characters), AWCV-100K, captured by general cameras, and propose a VCRec for sparse visual features and complex character structures.

Meiqi Wu, Yuzhong Zhao, Weiqiang Wang are with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China, E-mail: wumeiqi18, zhaoyuzhong20@mails.ucas.ac.cn, wqwang@ucas.ac.cn.

Yuanqiang Cai is currently a lecturer at the Beijing University of Posts and Telecommunications, Beijing 100876, China, E-mail: caiyuanqiang15@mails.ucas.ac.cn.

Shiyu Hu is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: hushiyu2019@ia.ac.cn.

Kaiqi Huang is with the Center for Research on Intelligent System and Engineering and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China. E-mail: kqhuang@nlpr.ia.ac.cn.

of real-world scenarios, air-writing systems necessitate the utilization of readily available sensors, such as RGB-based cameras, and the inclusion of frequently used words for effective communication.

However, conventional air-writing methods (Fig. 1 upper) [8], [9], [10], [11], [12], [13], [14], [15] relies on accurately captured handwritten trajectories by complex sensors (*e.g.*, Radar [1], Smart Watch [2], Leap Motion [3], EEGs [4], IMU [5]), which are inflexible and challenging to seamlessly integrate into popular smart devices. Moreover, the corpus of some works [16], [17] could not cover commonly used words for communications.

In order to overcome these limitations, we utilize common RGB cameras to record videos of handwritten gestures, enabling practical implementation in real-world scenarios. We create a video-based air-writing benchmark called AWCV-100K, which consists of a vast collection of 8.8 million video frames. This benchmark encompasses a comprehensive corpus, encompassing 3,755 Chinese characters from the GB1 set, representing 99.7% of characters commonly used in daily communication [18].

Furthermore, we propose a simple yet effective two-stage solution called the video-based character recognizer (VCRec) (Fig. 1 lower) to tackle this challenging task. The key to its success lies in leveraging sparse visual features, which are commonly found in real-world applications due to low frame rates. In the first stage, we introduce a fingertip feature extractor to condense the sparse visual features into fingertip features. In the second stage, VCRec adopts a spatial-temporal sequence module to model the character, capturing temporal information from fingertip movements. Concurrently, we employ a stroke graph attention network (StrokeGAT) to represent the spatial structure of Chinese characters, enhancing the utilization of sparse visual features.

Finally, we include comprehensive quantitative and qualitative evaluations on the AWCV-100K benchmark, comparing our approach to existing models in the field of video-based air-writing. Through extensive experimental analysis, we demonstrate that our approach outperforms conventional state-of-the-art (SOTA) methods, achieving a 4.92% improvement in recognition accuracy on the constructed AWCV-100K dataset.

In general, the main contributions of this paper can be summarized as follows:

- **AWCV-100K Introduction:** Presented the pioneering video-based air-writing dataset with comprehensive corpus, AWCV-100K, comprising 8.8 million frames and 3,755 Chinese characters, addressing limitations by utilizing general RGB cameras for real-world applications.
- **VCRec Model Proposition:** Introduced VCRec, a two-stage character recognition model leveraging sparse visual features, achieving improved accuracy in air-writing.
- **Performance Enhancement:** Through extensive analysis, demonstrated a significant 4.92% accuracy improvement over existing methods on the AWCV-100K, advancing effective human-computer interaction in real-world.

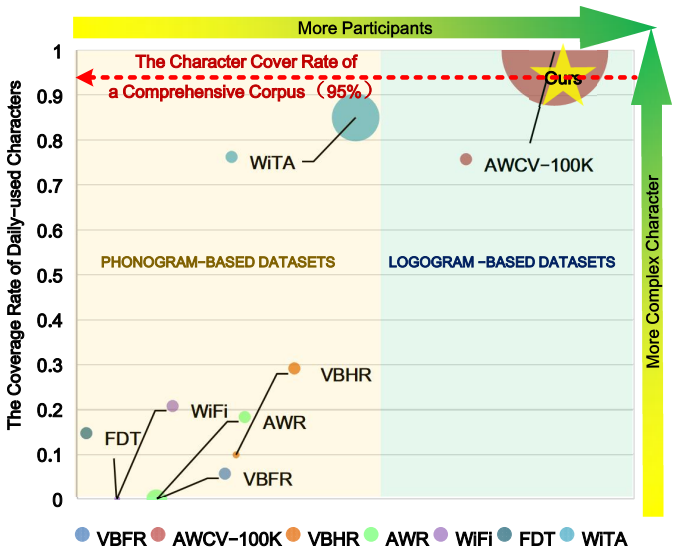


Fig. 2. Comparison between AWCV-100K with other benchmarks. Phonogram-based (*e.g.*, VBFR [19], VBHR [20], AWR [9], WiFi [21], FDT [16], WiTA [17]) and logogram-based benchmarks are selected for overall comparison. The bubble diameter is proportional to the total frames of the benchmark, and the vertical represents the coverage rate of daily-used characters in each benchmark. Obviously, the proposed AWCV-100K is the first logogram-based video dataset with a comprehensive corpus, more participants, and more complex characters.

II. Related Works

A. Air-Writing Datasets

Numerous air-writing systems have emerged as a new and intriguing research topic in the field of human-computer interaction, integrating various types of sensors in recent years. Depending on the type of sensor employed, air-writing datasets can be categorized into two main groups: trajectory-based datasets collected by complex sensors and video-based datasets collected by general cameras.

Trajectory-based Datasets. They were collected by complex sensors (*e.g.*, Radar [1]), data gloves [22], [23], [24], [25], hand motion sensors [26], [9], SmartWatch [2], Wifi [27], Leap Motion [3], EEGs[4], IMU [5], which acquired precise trajectories of air-writing. Kumar et al. [28] proposed a 3D English text air-writing system based on Leap Motion, collecting 560 sentences from 10 participants. However, most of them were not available. Qu et al. [13] introduced IAHCC-UCAS2016, a trajectory-based air-writing Chinese character dataset. Gan and Wang [29] proposed IAHEW-UCAS2016, a trajectory-based English word air-writing dataset. Gan et al. [30] proposed IAHCT-UCAS2018, a trajectory-based air-writing Chinese text dataset. These were public trajectory-based air-writing datasets collected by Leap Motion. With the introduction of public trajectory-based air-writing datasets, air-writing systems based on complex sensors have made some progress. However, due to their high cost and the difficulty of integrating these complex sensors into existing systems, this poses a challenge for their application in real-world scenarios.

Video-based Datasets. Video-based Datasets were collected by cameras (*e.g.*, Kinect [31], general RGB camera). Zhang et al. [8] presented a small video-based air-writing dataset that

Fig. 3. Examples of AWCV-100K. The figure shows a comparison of data under different lighting intensities and backgrounds. On the left of the figure are the video frames of datasets, and on the right are the labels of datasets. The blue circles represent overexposure due to strong illumination and the red box represents motion blur. **(TOP)** Data is collected under complex backgrounds and strong lighting conditions. **(BOTTOM)** Data is collected under simpler backgrounds and weaker lighting conditions.

utilized Kinect sensors. Unlike complex and expensive sensors, RGB cameras do not require physical contact, making them a convenient and cost-effective option. As shown in Fig. 2, we have summarized video-based air-writing datasets based on participant count, vocabulary categories, and the coverage rate of daily-used characters. VBFR [19], VBHR [20], FDT [16] had collected air-writing video datasets for English lowercase letters but have not accessed them. Kim et al. [17] introduced an English and Korean video-based air-writing dataset (WiTA), which was accessed. However, the corpora of the previous datsets did not achieve a coverage rate of over 95% for 'commonly used words', constraining research methodologies for real-world applications. Moreover, there hasn't been a video-based air-writing dataset focusing on logograms in the past, hindering general air-writing system developing.

B. Air-Writing Recognition Models

Most conventional air-writing recognition methods heavily rely on precise handwritten trajectories. For instance, Zhang et al. [8] utilized Kinect sensors for fingertip tracking, enabling controller-free motion tracking. Their system primarily focused on recognizing characters using the modified quadratic discriminant function (MQDF) classifier. In another approach, Kumar et al. [28] segmented each text into individual words and employed an LSTM-CTC structure for word recognition. Similarly, Gan and Wang [14] applied an LSTM-based sequence-to-sequence model for word recognition, achieving performance comparable to previous state-of-the-art methods. Gan et al. [32] revolutionized character representation by adopting skeleton graphs and introducing PyGT, a specialized transformer-convolutional network fusion. Furthermore, Wu et al. [33] introduced the attention convolutional loop network (ACRN), utilizing 1DCNN feature extraction followed by LSTM multi-head attention mechanism classification. Their experiments on CASIA-OLHWDB2.0-2.2 [34] and IAHCT-UCAS2018 [30] demonstrated higher recognition accuracy.

Recently, video-based air-writing was proposed by Kim et al. [17], introducing residual network architectures inspired by 3D ResNet. Additionally, Tan et al. [35] proposed transformer architectures for air-writing recognition. However, these methods did not particularly focus on the visual semantics and spatial features of fingertips, which led to lower performance.

C. Fingertip Detection and Tracking

In the realm of human-computer interaction (HCI), fingertip detection and tracking have been explored. Initially, Liang et al. [36] employed palm-to-hand outline distance measures for fingertip identification, further refined [37] using the hand's natural structure. However, challenges persisted with segmentation quality due to reliance on cues like color, depth, and motion [8]. Preceding these methods were model-based 3D gesture tracking approaches [38], [39], though these demanded significant computational resources and ample training data, limiting their real-time applicability. Contrastingly, MediaPipe, an open-source gesture recognition framework, offered real-time detection of 21 key point coordinates across various platforms, achieving impressive frame rates like 909 FPS on the iPhone 11, with a mean square error of 9.817mm [40], [41]. Recently, in the field of visual object tracking, researchers have devised various strategies such as dynamic attention-guided multi-trajectory methods [42], dynamic feature assignment frameworks like DFAT [43], occlusion-aware networks

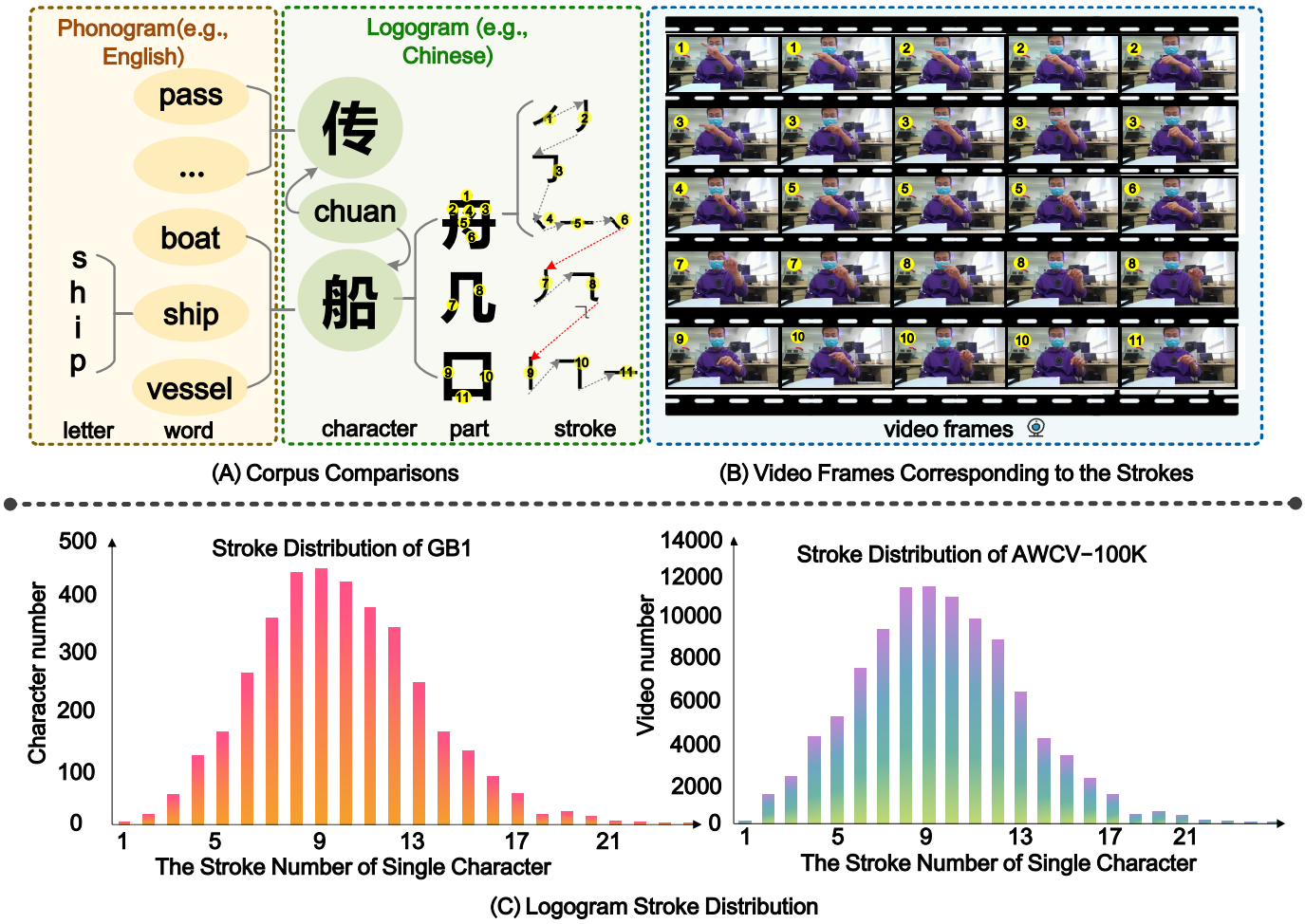


Fig. 4. More Complex and Comprehensive Corpus. The figure shows the characteristics of AWCV-100K. **(A)** Phonograms are composed of letters and easy to identify (*i.e.*, "`ship`" consists of the letters "`s,h,i,p`"). Logograms are described as having one pinyin corresponds to multiple characters. Each character is made up of parts, which can be further divided into strokes (*i.e.*, pinyin "`chuan`" with yellow numbers representing the stroke order). **(B)** Successive frames represent a stroke (*i.e.*, the first two frames of the video correspond to the first stroke in figure), which are collected by general camera under cluttered background and natural light. **(C)**The stroke distributions of GB1 and AWCV-100K respectively.

like SiamON [44], and the Siamese-based Twin Attention Network [45], alongside simplified long-term tracking [46], which could bolster gesture recognition and improve object manipulation by complementing existing fingertip detection and tracking methods.

III. AWCV-100K Dataset

Mainstream real-world devices typically utilize standard RGB cameras and necessitate comprehensive coverage of commonly used words for effective communication. To tackle these challenges, we introduce AWCV-100K, a pioneering video-based air-writing dataset designed for real-world applications. This extensive dataset is collected using general cameras and involves a large number of participants.

A. Data Collection

We have utilized an air-writing platform for data acquisition. Initially, we have briefed the participants on the data collection procedure. They have been instructed to assume that a perfect AI system would decode their air-writing, encouraging them to write as naturally as possible. Each participant has then composed approximately 500 words in Chinese, resulting in a total collection of 102,688 videos. During each data collection session, participants have had the flexibility to adjust the camera view, accommodating different angles and positions. The image sequences have been derived from real-time recorded videos at 30 frames per second (FPS). The examples of AWCV-100K are shown in Fig. 3.

B. Checkout Flow

We have implemented a stringent data review process to uphold the benchmark's quality. Trained professional collectors understand the nuances of the air-writing task and undertake preliminary work with a self-inspection process. Subsequently, verifiers conduct a second-round review of the collected data. Finally, authors make the ultimate judgment whether to accept or reject the data in the third-round confirmation. Any rejection during self-check, verification, or data acceptance necessitates recollection. We believe this three-round verification mechanism ensures the generation of a high-quality dataset.

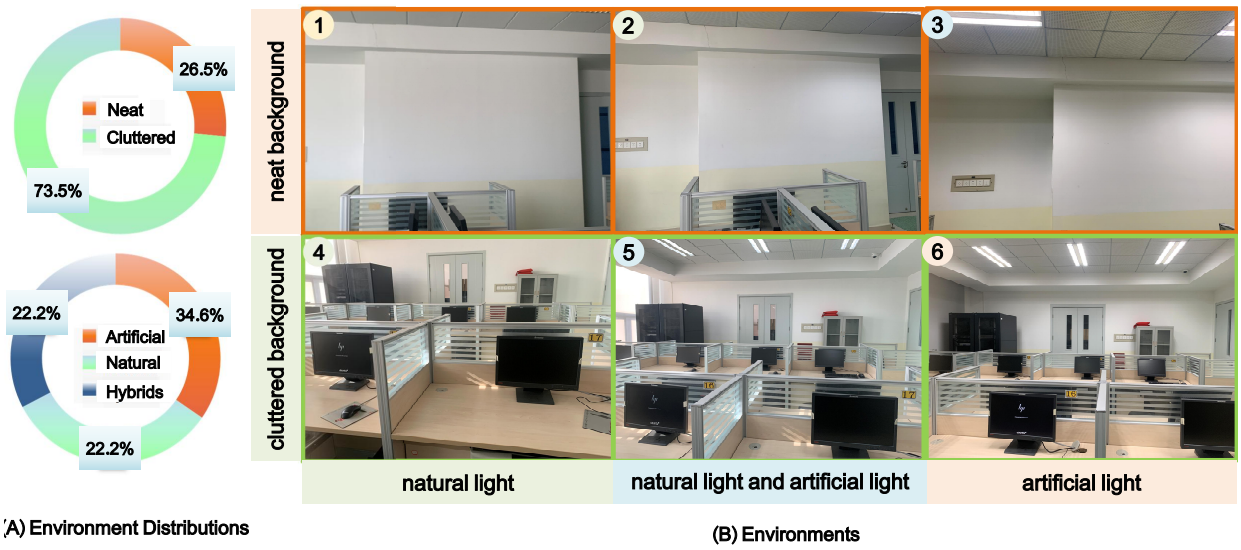


Fig. 5. Statistical Analysis Environments in AWCV-100K. (A) Depicts the diverse environmental distributions within AWCV-100K. (B) This section illustrates the environmental combinations of two backgrounds (*i.e.*, neat background and cluttered background) and three types of light (*i.e.*, natural light, artificial light, and their hybrids).

TABLE I
SUMMARY OF THE PARTICIPANT INFORMATION STATISTICS. STATISTICS ARE GATHERED ON PARTICIPANTS' AGE, GENDER, HANDEDNESS, AND HAND SIZE.

Metric	Type	Value
Gender	Male	145/211
	Female	66/211
	Neutral	0/211
Age	Range	20 - 30
	Average	23.33
	s.t.d.	1.39
Comfort-Hand	Left	1/211
	Right	210/211
	Both	0/211
Hand Width	[6cm, 8cm]	67/211
	(8cm, 10cm]	140/211
	$>$ 10cm	24/211
Hand Length	[15cm, 17cm]	15/211
	(17cm, 19cm]	60/211
	(19cm, 20cm]	131/211
	$>$ 20cm	5/211

C. Challenge Attributes

Sparse Visual Features. In real-world scenarios, mainstream devices only incorporate standard RGB cameras. In order to be more widely applicable in the real-world, we collect datasets by general RGB cameras, as illustrated in Fig. 4 (B). However, compared to sophisticated sensors (such as Leap Motion, which operates at 120FPS), general cameras often capture fewer frames per second ($e.g.$, 30FPS), resulting in the acquisition of relatively sparse features.

More complex corpus. Phonograms usually consist of only a few dozen letters ($e.g.$, English has 26 letters, German has 27, Russian has 33, *etc.*). Logograms, on the other hand, comprise thousands of characters ($e.g.$, Chinese has 3,755 characters only in the GB1 set, *etc.*). However, previous video-based air-writing datasets focused on phonograms and did not include a comprehensive corpus (a corpus vocabulary size achieving a coverage rate of over 95% for 'commonly used words'). To bridge this gap, we have constructed AWCV-100K with a comprehensive logogram corpus. It includes 3,755 Chinese characters from the GB1 set, encompassing 99.7% of characters used in daily communication [18], forming a comprehensive corpus. As shown in Fig. 4 (A), Chinese characters have complex structures, which are composed of multiple parts, each consisting of many strokes. The stroke distributions of GB1 and AWCV-100K are shown in Fig. 4 (C), which refers the stroke number of character is various. Moreover, Chinese characters are formed in various ways, resulting in highly intricate structures. These factors have presented greater challenges for video-based air-writing recognition.

More Various Environments. As shown in Fig. 5, we have collected data in various environments, encompassing different lighting conditions ($e.g.$, artificial light, natural light, and a combination of both) and backgrounds ($e.g.$, a neat background and a cluttered background) to ensure robustness against real-world scenario variations. Additionally, during data collection, we specifically have focused on capturing data under diverse weather conditions and at different times of the day (morning, afternoon, evening) to more accurately simulate natural light variations in real-world environments. Furthermore, participants have been allowed to adjust their seats, cameras, *etc.* We have varied viewpoints (camera distance, angle, and position) during different data collection processes to enhance the diversity of AWCV-100K.

More Participants. As depicted in Table I, which summarizes participant statistics, we have recruited a total of 211 individuals (male: 145, female: 66), who are native Chinese speakers proficient in both reading and writing. Participants'

TABLE II
COMPARISON OF DATASETS. THE PROPOSED AWCV-100K DATASET IS THE MOST COMPREHENSIVE AND PROVIDES RICH TYPES OF DATA INSTANCES. OUR DATASET SUPPLIES VIDEOS CONTAINING SEMANTIC TEXT WRITTEN IN THE AIR, WHICH CAPTURES THE INTERDEPENDENCE BETWEEN GESTURES FOR DIFFERENT CHARACTERS. (CRD, SEM, K, E, C, AND N IN THE TABLE STAND FOR INCLUSION OF COVERAGE RATE OF DAILY-USED CHARACTERS, SEMANTIC WORDS, KOREAN, ENGLISH, CHINESE, AND NUMBERS, RESPECTIVELY.)

Dataset	Year	People	Frames	CRD	Sem	Language	Sensor	Illumination	Access
VBFR [19]	2007	69	-	-	-	E	RGB	-	-
VBHR [20]	2012	21	-	-	-	E	RGB	-	-
ANWE [8]	2013	-	44,522	-	-	ECN	Depth	-	-
AWR [9]	2015	22	-	-	✓	E	Motion	-	-
PGEI [10]	2016	24	93,729	-	-	EC	Depth	-	-
WiFi [21]	2018	5	-	-	-	E	WiFi	-	-
FDT [16]	2019	5	-	-	-	EN	RGB	-	-
WiTA [17]	2021	122	1,757,307	89.9%	✓	EK	RGB	-	✓
AWCV-100K (Ours)	2023	**211**	**8,819,068**	**99.7%**	✓	C	RGB	✓	✓

hand lengths have ranged from 15.7 cm to 22.5 cm (M=18.29 cm, SD=1.05 cm), and their hand widths have ranged from 6.1 cm to 10.5 cm (M=8.16 cm, SD=0.73 cm). The diverse hand sizes and writing styles among numerous participants pose challenges to the accuracy of video-based air-writing recognition.

Others. As depicted in Fig. 3, rapid fingertip movements resulting in motion blur and problems arising from excessive illumination create hurdles in acquiring precise trajectory features. These issues specifically affect the image's sharpness and contours, adding complexity to the recognition process.

D. Dataset Comparison.

Table II presents a summary of the air-writing datasets collected in this study, comparing them with previous studies and highlighting the significant advantages of our dataset. ***(1) Originality:*** Our dataset stands as the sole publicly accessible study focusing on logograms, offering invaluable support for research into video-based air-writing in real-world scenarios. ***(2) Comprehensiveness:*** Our dataset achieves comprehensive coverage of the Chinese corpus by encompassing all characters of the GB1 set. In contrast, other datasets primarily concentrate on phonograms and provide limited coverage via select word videos, posing challenges in real-world applications. ***(3) Authenticity and Diversity:*** Our dataset spans a wide array of acquisition dimensions necessary for real-world applications, including scenes captured with general cameras. Additionally, we encompass diverse environments and various illuminations to closely mimic realistic scenarios. Moreover, by incorporating participants with diverse hand sizes and writing styles, our dataset mirrors the diversity among users. These elements present significant challenges while providing researchers with invaluable insights to study algorithm robustness in real-world scenarios.

E. Evaluation Protocol

To measure the recognition performance, both the correct rate (CR) and the accurate rate (AR) are used as the performance metrics, defined in the ICDAR 2013 Chinese handwriting recognition competition [47]. Specifically,

$$CR = (N_t - D_e - S_e)/N_t, \tag{1}$$

$$AR = (N_t - D_e - S_e - I_e)/N_t, \tag{2}$$

where N_t is the total number of characters in the test ground-truth sentences, while D_e, S_e, I_e denote deletion error, substitution error, and insertion error, respectively, between predictions and test ground-truth sentences. Additionally, in this paper, the sequence length of characters is set to 1.

IV. METHODOLOGY

A. Overview

As shown in Fig. 6, we propose the Video-based Character Recognizer (VCRec), a two-stage method that is simple yet effective in addressing sparse visual features. The video initially undergoes adaptive fingertip feature extraction via the fingertip feature extractor (Sec. IV-B). Subsequently, these adaptive fingertip features are processed within the spatio-temporal sequence module (Sec. IV-C), where they are encoded to extract intrinsic characteristics along two distinctive dimensions. More specifically, the temporal feature encoder captures temporal aspects of the fingertip, while the spatial feature encoder handles the fingertip's spatial features. Finally, the decoder decodes the fusion features into the character.

B. Fingertip Feature Extractor

Due to the low frame rate in real-world scenarios, the key lies in utilizing sparse visual features. To address this challenge, we introduce a Fingertip Feature Extractor to compress sparse visual features into fingertip features. As shown in Fig. 7, the video is first inputted by the Fingertip Tracker [40], [41] to obtain fingertip trajectories and encoded into fingertip features by Fingertip Representation.

We represent each trajectory with its derivatives including the offsets of positions and the writing directions rather than its raw absolute coordinates, which can effectively describe the next movement of strokes. As shown in Fig. 7, the following representations are calculated for the t-th point (p_t, q_t, s_t) of the trajectory: (1) the offsets of XY-coordinates, Δp and Δq; (2) the cosine and sine of the writing direction α; (3) the cosine and sine of the curvature β; (4) the change of the stroke identity. As a result, each point (p_t, q_t, s_t) is represented as an eight-dimensional vector $\mathbf{x}_t$ at time step $t, t \in \mathbb{Z}$, $i.e.$,

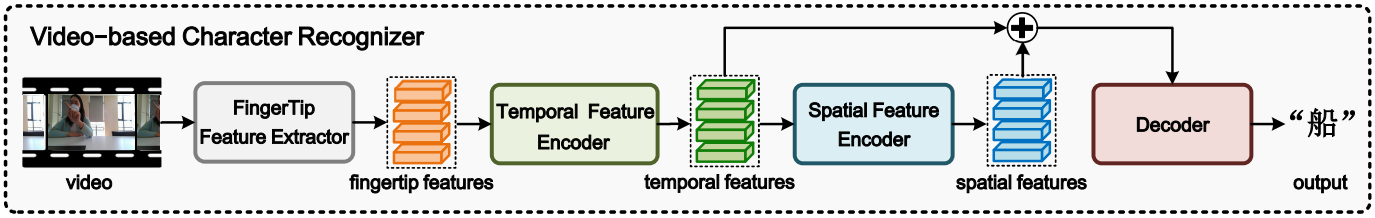


Fig. 6. Overview of the Video-based Character Recognizer (VCRec). The VCRec comprises a Fingertip Feature Extractor, a Spatio-Temporal Sequence Module that encompasses both Temporal Feature Encoder and Spatial Feature Encoder, as well as a CTC Decoder.

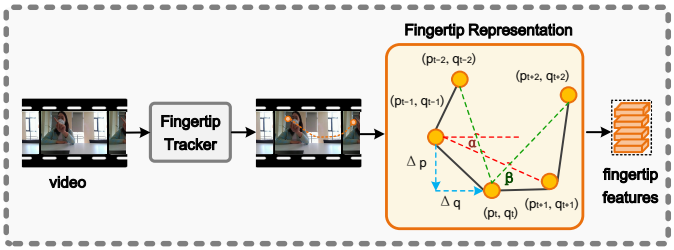


Fig. 7. Fingertip Feature Extractor. The video is first inputted by the Fingertip Tracker to obtain fingertip trajectories and encoded into fingertip features by Fingertip Representation.

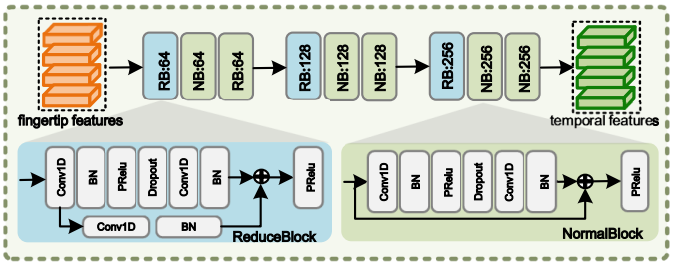


Fig. 8. Temporal Feature Encoder. It is constructed as a hierarchical framework, which is primarily composed of ReduceBlock and NormalBlock.

$$\mathbf{x}_t = [\Delta p_t, \Delta q_t, \sin\alpha, \cos\alpha, \sin\beta, \cos\beta, \mathbb{I}(s_t = s_{t+1}), \mathbb{I}(s_t \neq s_{t+1})]. \tag{3}$$

C. Spatio-Temporal Sequence Module

The Spatio-Temporal Sequence Module encodes the fingertip features of the character into probabilities as:

$$\mathbf{p} = f_c(f_g(f_r(\mathbf{x})), f_r(\mathbf{x})), \tag{4}$$

where $\mathbf{x}$ is the fingertip features predicted by the Fingertip Feature Extractor, f_r is the Temporal Feature Encoder, f_g is the Spatial Feature Encoder, and f_c is a decoder head that maps the deep feature into the character probabilities $\mathbf{p}$. During training, the cross-entropy loss supervises $\mathbf{p}$. During inference, the character with the highest probability is selected as the prediction.

Temporal Feature Encoder is uniquely tailored for encoding the temporal information of the fingertip feature sequences, utilizing mainly 1D convolution operators. As shown in Fig. 8, Temporal Feature Encoder is constructed as a hierarchical framework, which is primarily composed of ReduceBlock and NormalBlock inspired by Gan et al. [14]. Both types of conv-blocks utilize the techniques, like the residual connection [48], batch normalization [49], dropout [50], and parametric rectified linear unit (PReLU) [51], to speed up the network training and also address the over-fitting problem. The ReduceBlock, contains an extra convolution branch to adopt the residual connection when the numbers of input and output channels are different; if the convolution stride is set to 2, the ReduceBlock will downsample the sequence over the time dimension to increase the receptive field of the convolution.

Through Temporal Feature Encoder, the fingertip feature sequence is encoded by $\mathbf{Z} = f_r(\mathbf{x})$, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l]^{\mathrm{T}}, \mathbf{Z} \in \mathbb{R}^{l\times c}$, where l represents the length of feature in the time dimension and $\{\mathbf{z}_i\}_{i\in 1,2\cdots,l}$ are the features of different clips of the fingertip trajectory.

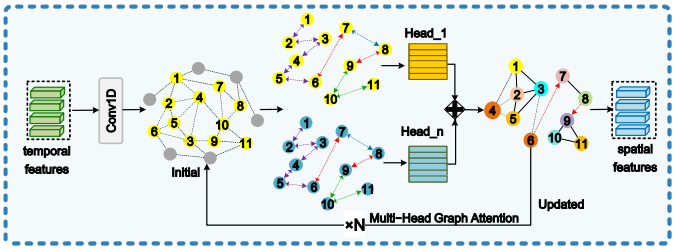


Fig. 9. Spatial Feature Encoder. It models the spatial structure of the character through a graph attention network, we have named it **StrokeGAT**.

Spatial Feature Encoder is used to encode the spatial information of the fingertip feature sequences based on feature $\mathbf{Z}$. As shown in Fig. 9, by treating the features of different clips of the trajectory (*i.e.*, $\{\mathbf{z}_i\}_{i\in 1,2\cdots,l}$) as the graph nodes, the Spatial feature encoder models the spatial structure of the character through a graph attention network [52], we have named it **StrokeGAT**. Through StrokeGAT, the temporal feature $\mathbf{Z}$ is encoded by $\overline{\mathbf{Z}} = f_g(\mathbf{Z})$, where $\overline{\mathbf{Z}} = [\overline{\mathbf{z}}_1, \overline{\mathbf{z}}_2, \cdots, \overline{\mathbf{z}}_l]^{\mathrm{T}}, \overline{\mathbf{Z}} \in \mathbb{R}^{l\times c}$.

Specifically, the decoder head is used to add the temporal features and spatial features of the fingertip movement trajectory and map them into probabilities of characters. To achieve a faster inference speed, we simply add the two types of features and then utilize a single-layer fully connected layer to map the feature to $\mathbf{p}$, *i.e.*, $\mathbf{p} = f_c(\overline{\mathbf{Z}} \oplus \mathbf{Z}) = \mathrm{FC}(\frac{1}{l}\sum_{i=1}^{l}[\overline{\mathbf{z}}_i^{\mathrm{T}} \oplus \mathbf{z}_i^{\mathrm{T}}])$, where FC and $[\oplus]$ denote the fully connected layer and add operator respectively.

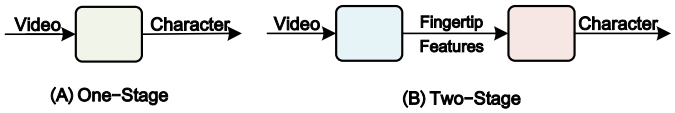


Fig. 10. Architectures of Air-Writing Recognition Method. (A) The one-stage method refers to encoding the features of the video and then decoding the characters. (B) The two-stage method refers to first extracting the fingertip trajectories, and then encoding the fingertip trajectories. Finally, decode the characters.

V. Experiments

A. Datasets

Our proposed baseline method VCRec underwent comparisons across the following datasets: the video-based air-writing datasets AWCV-100K (Ours), WiTA [17], and the trajectory-based air-writing datasets IAHCC-UCAS2016 [12] and IAHEW-UCAS2016 [29].

AWCV-100K is a video-based air-writing Chinese character dataset, including 102,688 videos and comprising a total of 8.8 million video frames. The benchmark includes 3,755 Chinese characters from the GB1 set, encompassing 99.7% of characters used in daily communication, forming a comprehensive corpus. To ensure the developed model is robust to variations among different individuals, we have partitioned the dataset into three sets (*i.e.*, training, validation, and testing) in an approximate ratio of 8:1:1, dividing the data by person.

WiTA [17] is a video-based air-writing photogram dataset. Only the English portion was used in this experiment, which included 10,620 video sequences from 122 participants. The data were sourced from an RGB camera with a frame rate of 29 fps, and all video frames were converted to 224×224 pixel images.

IAHCC-UCAS2016 [12] is a public trajectory-based air-writing Chinese character dataset, where each character is written in the midair within a single stroke. The dataset contains 431,825 samples of 3,755 different Chinese characters.

IAHEW-UCAS2016 [29] is a public large-vocabulary trajectory-based air-writing English word dataset. The dataset is contributed by 324 different participants and contains 150,480 recordings covering 2,280 English words.

B. Implementation Details

The experiment utilizes a Spatio-Temporal Sequence Module with a dropout probability of 0.2 in each conv-block and fully connected layer to ensure generalization. As shown in Fig. 9, N is 1, that is, one graph attention layer is used. n is 8, that is, using the 8-head self-attention mechanism. The proposed architecture is implemented using PyTorch [53] and initialized with default parameters after resizing images to 112×112. Optimization is achieved using the Adam [54] algorithm with a mini-batch size of 8. The initial learning rate is set to 0.001 and is decreased by a factor of 0.01 when recognition performance plateaus. The experiments are conducted on 4 NVIDIA TITAN RTX 24G GPUs.

TABLE III
Results of Different Methods on AWCV-100K.

	Method	Architecture	AR(%) ↑	Params(M) ↓	FPS ↑
One Stage	**CNN+LSTM** [55]	CNN/RNN	3.31	43.2	20.2
	TwoStream [55]		5.64	62.6	48.7
	C3D [55]		4.71	93.3	30.3
	ST-MC [17]		14.25	17.5	272.9
	ST-rMC [17]		16.02	52.4	306.7
	ST-R(2+1)D [17]		4.51	52.4	126.3
	ST-R3D [17]		23.40	55.4	168.7
	ViT [56]	Transformer	21.51	86.7	16.9
Two Stage	**VCRec (Ours)**	CNN+GAT	**52.43**	3.9	88.2

TABLE IV
Results of Different ST Methods on AWCV-100K. S represents the Spatial Feature Encoder and T represents the Temporal Feature Encoder.

Method	ST	AR(%) ↑	Params(M) ↓
1D-TCRN [30]	**T**	45.31	5.6
LSTM [57]	**T**	43.23	6.5
1DCNN [58]	**T**	47.51	1.1
Transformer [59]	**T**	40.11	36.2
VCRec (Ours)	**ST**	**52.43**	3.9

C. Results and Analysis

Comparison of Different Architectures. As shown in Fig. 10, methods are categorized into two groups based on their utilization of fingertip features: one-stage (not using fingertip features) and two-stage (using fingertip features). For video-based air-writing recognition tasks, one-stage methods ($e.g.$, ST-R3D, ST-(2+1)D [17]) encode the video frames directly for character recognition. To address sparse visual features, we first propose the two-stage architecture, which refers to extracting the fingertip features from sparse visual features and then encoding the fingertip features for character recognition.

Table III illustrates the performance of different model architectures on the AWCV-100K dataset. For the one-stage architecture, aside from ST-R3D [17], we have conducted experiments with various classic architectures ($e.g.$, C3D [55], TwoStream [55] and so on) used for modeling video sequences. Additionally, we have also utilized the ViT architecture. Several insights can be derived from these findings. Our proposed two-stage method significantly outperforms the one-stage approach, highlighting the crucial importance of fingertip features in air-writing recognition. Compared with ST-R3D [35], the previous SOTA method in video-based air-writing, VCRec (Ours) enhances accuracy by 29.03% on the AWCV-100K dataset. ViT architecture has been not very effective, possibly because the transformer architecture emphasizes global features but struggles to focus on specific visual characteristics. In the case of air-writing, effectively capturing visual features is particularly challenging due to their sparsity, thus leading to difficulty in achieving optimal performance.

Comparison of Different Spatio-Temporal Sequence Modules. As shown in Fig. 11, we design different Spatio-Temporal Sequence Module architectures. Fig. 11 (A) is the first structure without a spatial feature encoder. Fig. 11 (D) is the structure of VCRec (Ours). We conduct some experiments

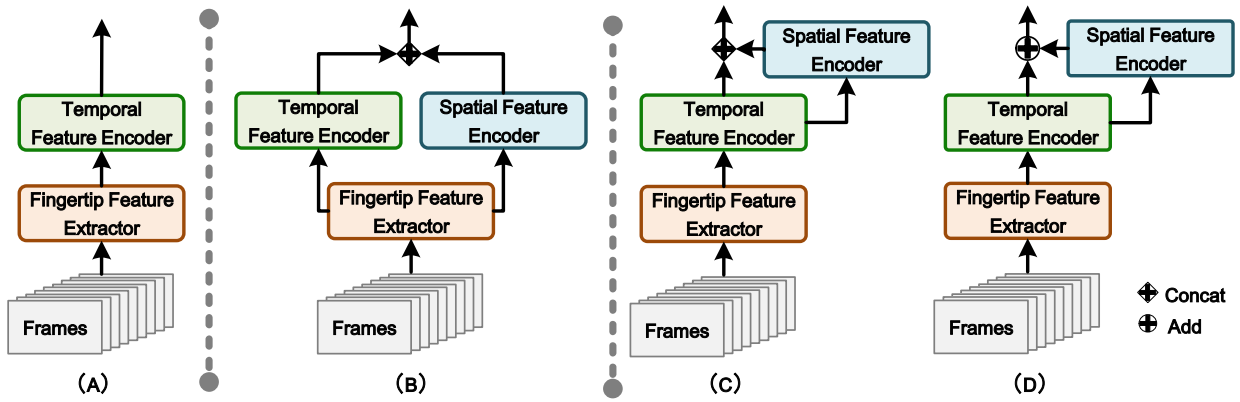


Fig. 11. Feature Fusion Strategy. (B), (C) and (D) are three different architectures of feature fusion strategies. (A) is the model without a Spatial Feature Encoder.

TABLE V
RESULTS OF DIFFERENT FEATURE FUSION STRATEGIES ON AWCV-100K. A, B, C, AND D CORRESPOND TO THE FOUR DIFFERENT STRUCTURES IN FIG. 11.

Feature Fusion Strategy	AR(%) ↑	Params(M) ↓
A	47.51	1.1
B	48.19	3.7
C	51.77	3.9
D(Ours)	**52.43**	3.9

shown in Table IV. The Transformer model we're using has 2 heads. The character recognition accuracy of the VCRec (Ours), is significantly better than that of the other two-stage methods, which proves that the spatial structure of the character is very important. Our preliminary exploration of modeling the spatial structure of characters has achieved good results. Compared with the temporal model, VCRec (Ours) has a 4.92% performance improvement.

Moreover, we have designed three different spatio-temporal feature fusion strategies, depicted in Fig. 11 as (B), (C), and (D). We conduct the relevant experiments for these feature fusion strategies on AWCV-100K, and the results are shown in Table V. Among them, comparing A, B, C and D have better accuracy, which shows the importance of modeling character structures. Comparing B, C and D behave better, which shows that higher-level features are more effective when modeling character structures by spatial feature encoder. Compared with C and D, D achieves the SOTA, which shows that it is more effective to apply structural information directly to stroke features.

Visual Analysis of VCRec. The analysis of Fig. 12 unveils a compelling relationship between the intricacy of characters and the corresponding response within the StrokeGAT feature map. As depicted in the lower section of the figure, characters exhibiting higher complexity appear to induce a more pronounced and intensified response in the StrokeGAT feature map located in the upper part.

This correlation signifies StrokeGAT's remarkable ability to capture and represent the intricate web of stroke connections

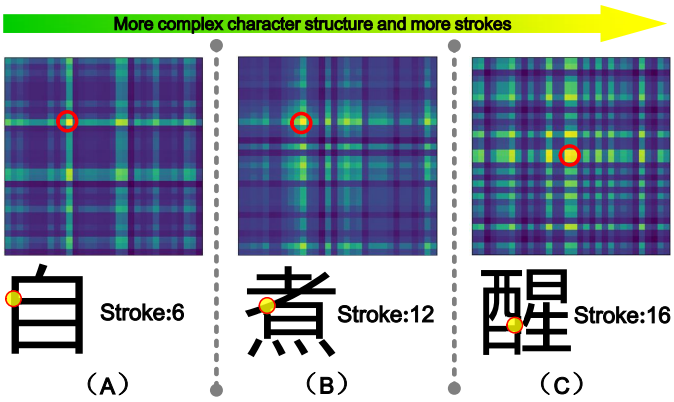


Fig. 12. The Visualization of StrokeGAT. (A), (B) and (C) represent three different characters. From left to right, the character structure becomes more complex and the number of strokes increases. **(TOP)** The visualization of character features through the StrokeGAT, the brighter the color, the closer the dependency between the features. **(BOTTOM)** Character and its stroke count, where the circle part corresponds to the highlighted part in the visualization.

inherent in various characters. The heightened highlighting observed in the feature map suggests that StrokeGAT excels in discerning and emphasizing these intricate stroke patterns, providing a deeper insight into the structural composition of characters.

D. Ablation Study

Comparison of Different Temporal Feature Encoder. We conducted ablation studies on the Temporal Feature Encoder, as detailed in Table VI. In the 4th row, we observed a significant decline in performance when the Temporal Feature Encoder was not utilized, resulting in a 12.76% lower performance compared to our method. Comparing the experimental results in rows 1-3 and our method, it is evident that 1DCNN excels in modeling temporal sequences.

Character Structure Model. In our quest to address sparse visual features, our focus delved into character spatial structures, leading to the proposition of the Spatial Feature Encoder strokeGAT. Given the significant strides of graph convolutional networks (GCN) in handling diverse unstructured data, our

TABLE VI
RESULTS OF DIFFERENT TEMPORAL FEATURE ENCODER IN VCREC ON AWCV-100K.

Temporal Feature Encoder	Spatial Feature Encoder	AR(%) ↑	Params(M) ↓
1D-TCRN [30]	StrokeGAT	49.31	8.6
LSTM [57]	StrokeGAT	47.22	9.5
Transformer [58]	StrokeGAT	44.11	39.2
✗	StrokeGAT	39.67	2.9
1DCNN(Ours)	StrokeGAT	**52.43**	3.9

TABLE VII
RESULTS OF DIFFERENT CHARACTER STRUCTURE MODEL IN VCREC ON AWCV-100K.

Character Structure Model	AR(%) ↑	Params(M) ↓
GCN [61]	51.88	3.8
SparseGAT [52]	49.34	4.1
StrokeGAT(Ours)	**52.43**	3.9

study ventured into character spatial exploration using GCN, SparseGAT [52], and StrokeGAT. As elucidated in Table VII, StrokeGAT emerged as the optimal performer, highlighting the Graph Attention Network's (GAT) adeptness in capturing extensive spatial dependencies crucial for character analysis. This reinforces GAT's superiority in discerning and modeling intricate spatial structures pivotal for character feature understanding.

Comparison of Different Decoders. In our analysis of AWCV-100K, we scrutinized various decoder models. The Connectionist Temporal Classification (CTC) technique, pioneered by Graves et al. [60], empowers models to learn direct mappings from input to output sequences sans explicit alignment requirements. Our experimentation, outlined in Table VIII, encompassed exploring diverse configurations of fully connected decoders. Notably, the 2-layer fully connected (2-layer FC) decoder emerged as the best performer.

In contrast to CTC, the 2-layer FC decoder exhibited superior performance, particularly excelling in single-character recognition tasks. This disparity highlights the distinct advantage of the FC decoder architecture over CTC, especially in accurately identifying individual characters within sequences.

E. Performance on other Dataset.

We construct different experiments on different languages ($e.g.$, English and Chinese) and forms ($e.g.$, trajectory-based and video-based) of datasets.

Table IX displays the performance of VCRec on trajectory-based datasets, which have exhibited even better performance on the IAHCC-UCAS2016 (a trajectory-based Chinese character dataset).

Since the WiTA consists of English words containing multiple English characters, we have adopted the concept of CTC [60] in those experiments instead of the FC decoder mentioned in our method. Table X displays the performance of VCRec on a video-based English dataset, which has demonstrated comparable performance on the WiTA (English). Table XI displays the performance of VCRec on trajectory-based datasets, which have exhibited comparable performance on the IAHEW-UCAS2016 (a trajectory-based English word dataset).

The results of the experiments on these English datasets have demonstrated comparable performance in other languages, indicating that our method exhibits strong generalization across different languages. In addition, we have proceeded to analyze that our approach showcases considerable effectiveness in modeling spatial structures within the context of the Chinese language, leveraging its inherent structural traits. Nevertheless, its performance exhibits a relatively diminished impact in English contexts, primarily due to the heightened emphasis on temporal information within the language.

TABLE VIII
RESULTS OF DIFFERENT DECODERS IN VCREC ON AWCV-100K.

Decoder	AR(%) ↑	Params(M) ↓
CTC [60]	49.66	3.9
1-layer FC	50.43	3.8
3-layer FC	51.43	3.9
2-layer FC(Ours)	**52.43**	3.9

TABLE IX
VCREC PERFORMS ON THE TRAJECTORY-BASED CHINESE CHARACTER DATASET IAHCC-UCAS2016 [12].

Trajectory-based Dataset	Method	AR(%) ↑
IAHCC-UCAS2016	**LSTM** [62]	93.18
	ViT-B [63]	93.85
	ViT-L [63]	93.91
	1DCNN [58]	96.78
	VCRec (Ours)	**96.85**

TABLE X
VCREC PERFORMS ON VIDEO-BASED ENGLISH DATASETS. WITA [17] IS A VIDEO-BASED ENGLISH DATASET. CER IS CHARACTER ERROR RATE.

Video-based Dataset	Method	CER(%) ↓
WiTA	**ST-rMC** [17]	92.94
	ST-R(2+1)D [17]	87.51
	ST-R3D [17]	29.24
	TR-AWR [35]	29.86
	VCRec+CTC	30.12

TABLE XI
VCREC PERFORMS ON TRAJECTORY-BASED ENGLISH DATASETS. IAHEW-UCAS2016 [29] IS A TRAJECTORY-BASED ENGLISH DATASET. CAR IS CHARACTER ACCURACY RATE.

Tarjectory-based Dataset	Method	CAR(%) ↑
IAHEW-UCAS2016	**LSTM+CTC** [62]	97.13
	LSTM+Decoder [64]	96.86
	1DCNN+Decoder [58]	97.45
	VCRec (Ours)	**96.51**

VI. Conclusion

In this work, we have presented the AWCV-100K dataset, a video-based air-writing dataset designed for real-world scenarios. To address the challenges posed by AWCV-100K, we propose the VCRec method, a two-stage architecture. This method initially compresses sparse visual features into fingertip features and then models fingertip feature sequences using a spatial-temporal sequence module. This module captures temporal information from fingertip movements and represents the spatial structure of Chinese characters. VCRec achieves an accuracy of 52.43% on the AWCV-100K dataset.

We anticipate that our dataset and the baseline model will stimulate research in real-world applications. For example, applying air-writing in healthcare for hands-free control in sterile environments, utilizing it in AR and VR for immersive experiences, integrating it into education and training scenarios, enhancing accessibility for individuals with disabilities, and implementing gesture-based interfaces in industrial settings for improved safety and efficiency in manufacturing processes.

The dataset, toolkit, and experimental results will be released to further advance air-writing research.

References

[1] S. Ahmed, W. Kim, J. Park, and S. H. Cho, "Radar-based air-writing gesture recognition using a novel multistream cnn approach," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23869–23880, 2022.

[2] V. Chandel and A. Ghose, "Nntrak: A neural network approach towards calculating air-writing trajectories in real-time with a smartwatch," in *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 374–379, 2023.

[3] J. Gan, W. Wang, and K. Lu, "A unified cnn-rnn approach for in-air handwritten english word recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2018.

[4] A. Tripathi, A. Gupta, A. P. Prathosh, S. P. Muthukrishnan, and L. Kumar, "Neuroair: Deep learning framework for airwriting recognition from scalp-recorded neural signals," 2023.

[5] H. Zhang, L. Chen, Y. Zhang, R. Hu, C. He, Y. Tan, J. Zhang, *et al.*, "A wearable real-time character recognition system based on edge computing-enabled deep learning for air-writing," *Journal of Sensors*, vol. 2022, 2022.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," *arXiv preprint arXiv:2103.10360*, 2021.

[8] X. Zhang, Z. Ye, L. Jin, Z. Feng, and S. Xu, "A new writing experience: Finger writing in the air using a kinect sensor," *IEEE MultiMedia*, vol. 20, no. 4, pp. 85–93, 2013.

[9] M. Chen, G. AlRegib, and B.-H. Juang, "Air-writing recognition—part i: Modeling and recognition of characters, words, and connecting motions," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 403–413, 2015.

[10] Y. Huang, X. Liu, X. Zhang, and L. Jin, "A pointing gesture based egocentric interaction system: Dataset, approach and application," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 16–23, 2016.

[11] M. Arsalan, A. Santra, K. Bierzynski, and V. Issakov, "Air-writing with sparse network of radars using spatio-temporal learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8877–8884, IEEE, 2021.

[12] N. Xu, W. Wang, and X. Qu, "Recognition of in-air handwritten chinese character based on leap motion controller," in *Image and Graphics: 8th International Conference, ICIG 2015, Tianjin, China, August 13–16, 2015, Proceedings, Part III*, pp. 160–168, Springer, 2015.

[13] X. Qu, W. Wang, K. Lu, and J. Zhou, "Data augmentation and directional feature maps extraction for in-air handwritten chinese character recognition based on convolutional neural network," *Pattern recognition letters*, vol. 111, pp. 9–15, 2018.

[14] J. Gan and W. Wang, "In-air handwritten english word recognition using attention recurrent translator," *Neural Computing and Applications*, vol. 31, pp. 3155–3172, 2019.

[15] J. Gan, W. Wang, and K. Lu, "Compressing the cnn architecture for in-air handwritten chinese character recognition," *Pattern Recognition Letters*, vol. 129, pp. 190–197, 2020.

[16] S. Mukherjee, S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Fingertip detection and tracking for recognition of air-writing in videos," *Expert Systems with Applications*, vol. 136, pp. 217–229, 2019.

[17] U.-H. Kim, Y. Hwang, S.-K. Lee, and J.-H. Kim, "Writing in the air: Unconstrained text recognition from finger movement using spatio-temporal convolution," *IEEE Transactions on Artificial Intelligence*, 2022.

[18] Y. Liu, H. Shu, and P. Li, "Word naming and psycholinguistic norms: Chinese," *Behavior research methods*, vol. 39, no. 2, pp. 192–198, 2007.

[19] L. Jin, D. Yang, L.-X. Zhen, and J.-C. Huang, "A novel vision-based finger-writing character recognition system," *Journal of Circuits, Systems, and Computers*, vol. 16, no. 03, pp. 421–436, 2007.

[20] A. Schick, D. Morlock, C. Amma, T. Schultz, and R. Stiefelhagen, "Vision-based handwriting recognition for unrestricted text input in mid-air," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 217–220, 2012.

[21] Z. Fu, J. Xu, Z. Zhu, A. X. Liu, and X. Sun, "Writing in the air with wifi signals for virtual reality devices," *IEEE Transactions on Mobile Computing*, vol. 18, no. 2, pp. 473–484, 2018.

[22] S. S. Fels and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *IEEE transactions on Neural Networks*, vol. 4, no. 1, pp. 2–8, 1993.

[23] T. Költringer, P. Isokoski, and T. Grechenig, "Twostick: Writing with a game controller," in *Proceedings of Graphics Interface 2007*, pp. 103–110, 2007.

[24] C. Amma, D. Gehrig, and T. Schultz, "Airwriting recognition using wearable motion sensors," in *Proceedings of the 1st Augmented Human international Conference*, pp. 1–8, 2010.

[25] C. Amma, M. Georgi, and T. Schultz, "Airwriting: Hands-free mobile text input by spotting and continuous recognition of 3d-space handwriting with inertial sensors," in *2012 16th International Symposium on Wearable Computers*, pp. 52–59, IEEE, 2012.

[26] C. Amma, M. Georgi, and T. Schultz, "Airwriting: a wearable handwriting recognition system," *Personal and ubiquitous computing*, vol. 18, pp. 191–203, 2014.

[27] Z. Fu, J. Xu, Z. Zhu, A. X. Liu, and X. Sun, "Writing in the air with wifi signals for virtual reality devices," *IEEE Transactions on Mobile Computing*, vol. 18, p. 473–484, feb 2019.

[28] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, "Study of text segmentation and recognition using leap motion sensor," *IEEE Sensors Journal*, vol. 17, no. 5, pp. 1293–1301, 2016.

[29] J. Gan and W. Wang, "In-air handwritten english word recognition using attention recurrent translator," *Neural computing applications*, 2019.

[30] J. Gan, W. Wang, and K. Lu, "In-air handwritten chinese text recognition with temporal convolutional recurrent network," *Pattern Recognition*, vol. 97, p. 107025, 2020.

[31] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[32] J. Gan, Y. Chen, B. Hu, J. Leng, W. Wang, and X. Gao, "Characters as graphs: Interpretable handwritten chinese character recognition via pyramid graph transformer," *Pattern Recognition*, vol. 137, p. 109317, 2023.

[33] Z. Wu, X. Qu, J. Huang, and X. Wu, "In-air handwritten chinese text recognition with attention convolutional recurrent network," in *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part II*, pp. 695–707, Springer, 2023.

[34] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *2011 International Conference on Document Analysis and Recognition*, pp. 37–41, 2011.

[35] X. Tan, J. Tong, T. Matsumaru, V. Dutta, and X. He, "An end-to-end air writing recognition method based on transformer," *IEEE Access*, 2023.

[36] H. Liang, J. Yuan, and D. Thalmann, "3d fingertip and palm tracking in depth image sequences," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 785–788, 2012.

[37] P. Krejov and R. Bowden, "Multi-touchless: Real-time fingertip detection and tracking using geodesic maxima," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–7, IEEE, 2013.

[38] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1793–1805, 2011.

[39] H. J. Chang, G. Garcia-Hernando, D. Tang, and T.-K. Kim, "Spatio-temporal hough forest for efficient detection–localisation–recognition of fingerwriting in egocentric camera," *Computer Vision and Image Understanding*, vol. 148, pp. 87–96, 2016.

[40] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[41] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[42] X. Wang, Z. Chen, J. Tang, B. Luo, Y. Wang, Y. Tian, and F. Wu, "Dynamic attention guided multi-trajectory analysis for single object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4895–4908, 2021.

[43] W. Zhang, L. Jiao, F. Liu, S. Yang, and J. Liu, "DFAT: dynamic feature-adaptive tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 43–58, 2023.

[44] C. Fan, H. Yu, Y. Huang, C. Shan, L. Wang, and C. Li, "Siamon: Siamese occlusion-aware network for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 186–199, 2023.

[45] H. Bao, P. Shu, H. Zhang, and X. Liu, "Siamese-based twin attention network for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 847–860, 2023.

[46] X. Xu, J. Zhao, J. Wu, and F. Shen, "Switch and refine: A long-term tracking and segmentation framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1291–1304, 2023.

[47] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "Icdar 2013 chinese handwriting recognition competition," in *2013 12th international conference on document analysis and recognition*, pp. 1464–1470, IEEE, 2013.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, PMLR, 07–09 Jul 2015.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[52] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[53] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration," *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, vol. 6, no. 3, p. 67, 2017.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[55] W. Kay, J. Carreira, K. Simonyan, B. Zhang, and A. Zisserman, "The kinetics human action video dataset," 2017.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[57] L. Sun, T. Su, C. Liu, and R. Wang, "Deep lstm networks for online chinese handwriting recognition," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 271–276, 2016.

[58] J. Gan, W. Wang, and K. Lu, "A new perspective: Recognizing online handwritten chinese characters via 1-dimensional cnn," *Information Sciences*, vol. 478, pp. 375–390, 2019.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

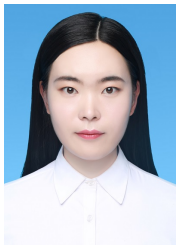
[60] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

[61] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[62] X.-Y. Zhang, F. Yin, Y.-M. Zhang, C.-L. Liu, and Y. Bengio, "Drawing and recognizing chinese characters with recurrent neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 849–862, 2018.

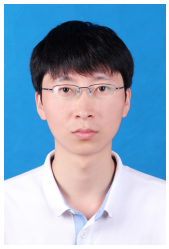
[63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[64] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.



Meiqi Wu received the MSc Degree from the University of Science and Technology of China In 2021 and continued to study for her doctorate at the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China, in the same year. Her current research interests include computer vision, human-computer interaction, and machine learning.

Kaiqi Huang received the BSc and MSc degrees from the Nanjing University of Science Technology, China, and the PhD degree from Southeast University. He is currently a full professor with the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences. He is also with the University of Chinese Academy of Sciences (UCAS), and the CAS Center for Excellence in Brain Science and Intelligence Technology. He has authored or co-authored more than 210 papers in important international journals and conferences, such as the IEEE TPAMI, IJCV, T-IP, T-SMCB, TCSVT, Pattern Recognition, CVIU, ICCV, ECCV, CVPR, ICIP, and ICPR. His current research interests include computer vision, pattern recognition, and game theory, including object recognition, video analysis, and visual surveillance. He is the co-chair and program committee member of more than 40 international conferences, such as ICCV, CVPR, ECCV, and the IEEE workshops on visual surveillance. He is an associate editor for IEEE Transactions on Systems, Man, and Cybernetics: Systems and Pattern Recognition.



Yuanqiang Cai received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently a lecturer at the Beijing University of Posts and Telecommunications. His research interests include object detection, multimedia content analysis, and text localization and recognition in images and videos. He has published more than 10 papers in referred conferences and journals including NeurIPS, AAAI, ACM MM, TCSVT, and PR.

Shiyu Hu , Ph.D., Institute of Automation, Chinese Academy of Sciences. Shiyu Hu received her PhD degree from the University of Chinese Academy of Sciences in Jan. 2024. She has authored or coauthored more than 10 research papers in the areas of computer vision and pattern recognition at international journals and conferences, including TPAMI, IJCV, NeurIPS, etc. Her research interests include computer vision, visual object tracking, and visual intelligence evaluation.

Yuzhong Zhao received a B.S. degree from Peking University, Beijing, China in 2020, and he is currently pursuing an M.E. degree at the University of Chinese Academy of Science, Beijing. His research interests include computer vision, scene text detection, and recognition.

Weiqiang Wang received the B.E. and M.S. degrees in computer science from Harbin Engineering University, in 1995 and 1998, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), China, in 2001. He is currently a Professor at the School of Computer Science and Technology, University of Sciences. His research interests include multimedia content analysis, computer vision, pattern recognition, and human-computer interaction.