Insertion Variant Calling in Long-Read Data Through a Transformer and Dual-Modal Fusion Model

Junjie Zhang, Chunxiao Lai, Jing Peng*, and Hao Wu
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan, China
jackjun@whut.edu.cn, pengjing@whut.edu.cn

***Abstract*—Insertions represent a significant class of structural variations (SV), and their accurate detection is critical for elucidating the mechanisms of genetic diseases. Although current deep learning methods based on long-read sequencing data have demonstrated significant promise in detecting insertion structural variations, most rely on single-modal feature representations—such as CIGAR images or numerical matrices—which fail to comprehensively capture the multi-dimensional signals of SV, thereby limiting detection accuracy and robustness. To address this limitation, we propose TDSV, a Transformer-based dual-modal fusion model for analyzing insertion structural variations in sequencing data. TDSV constructs an image-sequencing dual-modal feature representation by simultaneously extracting CIGAR images from candidate regions and multi-dimensional numerical tensors of sequence statistics, providing complementary information to the model. We further design a bidirectional cross-modal attention mechanism to enable deep interaction, alignment, and enhancement between the two modalities, allowing the model to synergistically capture local fine-grained patterns in images and global statistical regularities in sequences. To improve the detection of cross-region long insertion structural variations, a Transformer encoder is introduced to perform global modeling of the fused features, capturing long-range dependencies across consecutive sub-regions. Experiments on multiple real datasets show that TDSV significantly outperforms existing mainstream methods in several evaluation metrics, demonstrating particularly robust performance in low-coverage genomic regions.**

***Index Terms*—structural variant detection, image-sequence dual-modal fusion, bidirectional cross-modal attention, transformer model**

I. Introduction

Structural variations (SVs), defined as large-scale genomic alterations exceeding 50 base pairs [1], constitute a major source of human genetic diversity and play a critical role in the etiology of numerous genetic disorders [2]. These variations can disrupt gene structure and regulatory elements, precipitating severe clinical outcomes [3]. Specifically, insertion variants are clinically significant; for instance, expansions in the *FMR1* gene are the underlying cause of Fragile X syndrome, while insertions in the *HTT* gene lead to Huntington's disease [4]. Given their pathogenic potential, accurate detection of insertions is essential for elucidating disease mechanisms and advancing molecular diagnosis [5].

The advent of third-generation sequencing (TGS) technologies [6], characterized by long read lengths and uniform coverage [7], has significantly enhanced our capability to resolve large-scale structural variations [8], complex rearrangements, and repetitive genomic regions [9]. However, the inherent high error profiles of TGS data [10] pose a persistent challenge, increasing the risk of false positives in complex genomic regions. To mitigate these issues, numerous alignment-based heuristic algorithms have been developed, including cuteSV [11], SVIM [12], and DeBreak [13]. These methods infer SVs by analyzing hand-crafted alignment signatures in long reads. Nevertheless, their reliance on manually designed rules limits their adaptability to noise [14]. Sequencing artifacts or ambiguous alignments are frequently misinterpreted as genuine SV signals [15], leading to imprecise breakpoint localization and biased variant length estimation, especially in complex genomic contexts [16].

Recent advances in deep learning (DL) have shown considerable promise for structural variation detection. Unlike heuristic methods, DL approaches can autonomously learn intricate feature representations from large-scale labeled data, effectively capturing latent SV patterns. Several studies exemplify this potential [17]: SVcnn [18] utilizes convolutional neural networks to detect indels from image-encoded alignments; INSnet [19] applies recurrent neural networks to two-dimensional feature matrices for insertion detection; and DASV [20] employs dual-channel attention to combine image and base-level features for deletion calling. These methodologies strongly substantiate the utility of data-driven approaches in genomic variant calling.

Despite these advancements, deep learning-based SV callers face unresolved challenges, particularly regarding feature representation. Most existing models rely on single-modality inputs, which inevitably leads to information loss. For instance, converting alignment data solely into CIGAR-based images may obscure critical nucleotide-level sequence details, whereas compressing reads into numerical feature vectors often discards the two-dimensional spatial context inherent in alignment pileups. This loss of structural and sequential information is detrimental when resolving complex or closely

spaced SVs. Consequently, single-modal methods fail to fully encapsulate the multi-dimensional nature of variant signals, leaving the full potential of deep learning underexplored. To address these limitations, we present TDSV, a Transformer-based dual-modal framework specialized for detecting insertions from long-read sequencing data. Unlike existing methods that rely on static feature concatenation, TDSV integrates local CIGAR image representations with global sequence statistics through a dynamic fusion strategy. Our primary contributions are as follows:

- We introduce a comprehensive multi-modal feature encoding strategy that represents candidate SV regions using both CIGAR images and numerical feature tensors. This approach preserves both spatial read distributions and detailed sequence statistics, providing complementary morphological and statistical descriptors.
- We propose a bidirectional cross-modal attention network with an adaptive gating mechanism that dynamically weights sequence and image modalities. By computing attention in both directions and employing a learnable gating function, the model adaptively fuses and balances morphological and statistical features, suppressing noise and enhancing discriminability.
- We employ a Transformer encoder to perform global modeling on the fused feature representations. This architecture captures long-range dependencies across consecutive sub-regions, which is essential for accurately characterizing the structural complexity of large insertion variants.

II. Method

TDSV is a novel Transformer-based multimodal framework designed for high-precision insertion variant detection. It seamlessly integrates sequence statistics with CIGAR-derived visual representations. The framework processes aligned long-read data by simultaneously extracting numerical feature matrices and their corresponding CIGAR images. These inputs are encoded via depthwise separable convolutions to efficiently capture local variant patterns. To facilitate deep feature interaction, a bidirectional cross-modal attention mechanism with an adaptive gating module dynamically fuses these modalities, enabling the mutual enhancement of statistical and morphological features. Subsequently, the fused representations from consecutive genomic windows are processed by a Transformer encoder to model long-range dependencies, which is critical for resolving large-span long insertion structural variations. Finally, a post-processing clustering step refines variant positions and lengths, with the final predictions generated in standard VCF format.The overall architecture of the proposed TDSV framework is illustrated in Fig. 1.

A. Sequence Statistics Encoding

TDSV performs feature extraction based on alignment information by partitioning the reference genome into 200-bp non-overlapping subwindows. For each subwindow, it constructs a numerical feature matrix of size 200×5 by extracting five key sequence-derived features at every genomic site:

- **Read Depth**: The read coverage at the site.
- **Left Soft-Clip Count**: The count of left soft-clip breakpoints at the site.
- **Right Soft-Clip Count**: The count of right soft-clip breakpoints at the site.
- **Split Read Count**: The number of split alignments at the site.
- **Insertion Frequency**: The count of 'I' (insertion) operations derived from CIGAR strings.

Subwindows shorter than 200 bp are zero-padded to maintain consistent dimensions. A lightweight CNN with a Convolutional Block Attention Module (CBAM) encodes these features [21]. The CNN captures local spatial patterns within the feature matrix, while the CBAM enhances discriminative features by modeling channel and spatial dependencies, thereby suppressing less informative signals.

B. CIGAR Image Encoding

We implement an image encoding scheme based on CIGAR operations to visually represent structural variant regions. For each subwindow, we parse CIGAR strings from all overlapping reads and classify them into four core operations (M/=, I, D/N, S). These operations are mapped to specific RGB channels: insertions to [1,0,0] (red), matches to [0,1,0] (green), deletions to [0,0,1] (blue), and soft clips to [0,0,0] (black), set against a white [1,1,1] background. In the generated image, each row represents a read and each column corresponds to a reference site. The resulting images are resampled to a uniform $200 \times 200 \times 3$ resolution using nearest-neighbor interpolation to preserve discrete CIGAR boundaries. This encoding produces pseudo-color images with distinct biological semantics, visually conveying the SV distribution and spatial characteristics within the aligned region. These images serve as interpretable visual inputs for the subsequent deep learning models.

We implement a lightweight convolutional neural network based on depthwise separable convolutions to encode these CIGAR images [22]. The network begins with a stem convolution layer (16 filters, 7×7, stride 2) that performs initial feature extraction while reducing spatial dimensions. Subsequently, four sequential depthwise separable blocks process the features. Each block consists of a depthwise convolution for spatial filtering followed by a pointwise convolution for channel mixing. Batch normalization and ReLU activations are applied after each convolution operation. The architecture progressively increases the number of feature channels (32, 64, 128, 256) while reducing spatial resolution through strided convolutions. Finally, global average pooling condenses the feature maps into a compact 256-dimensional vector, which is then projected to the target dimensionality. With fewer than 0.5M parameters, this efficient design facilitates the processing of image sequences while maintaining strong representational capacity for SV pattern recognition.

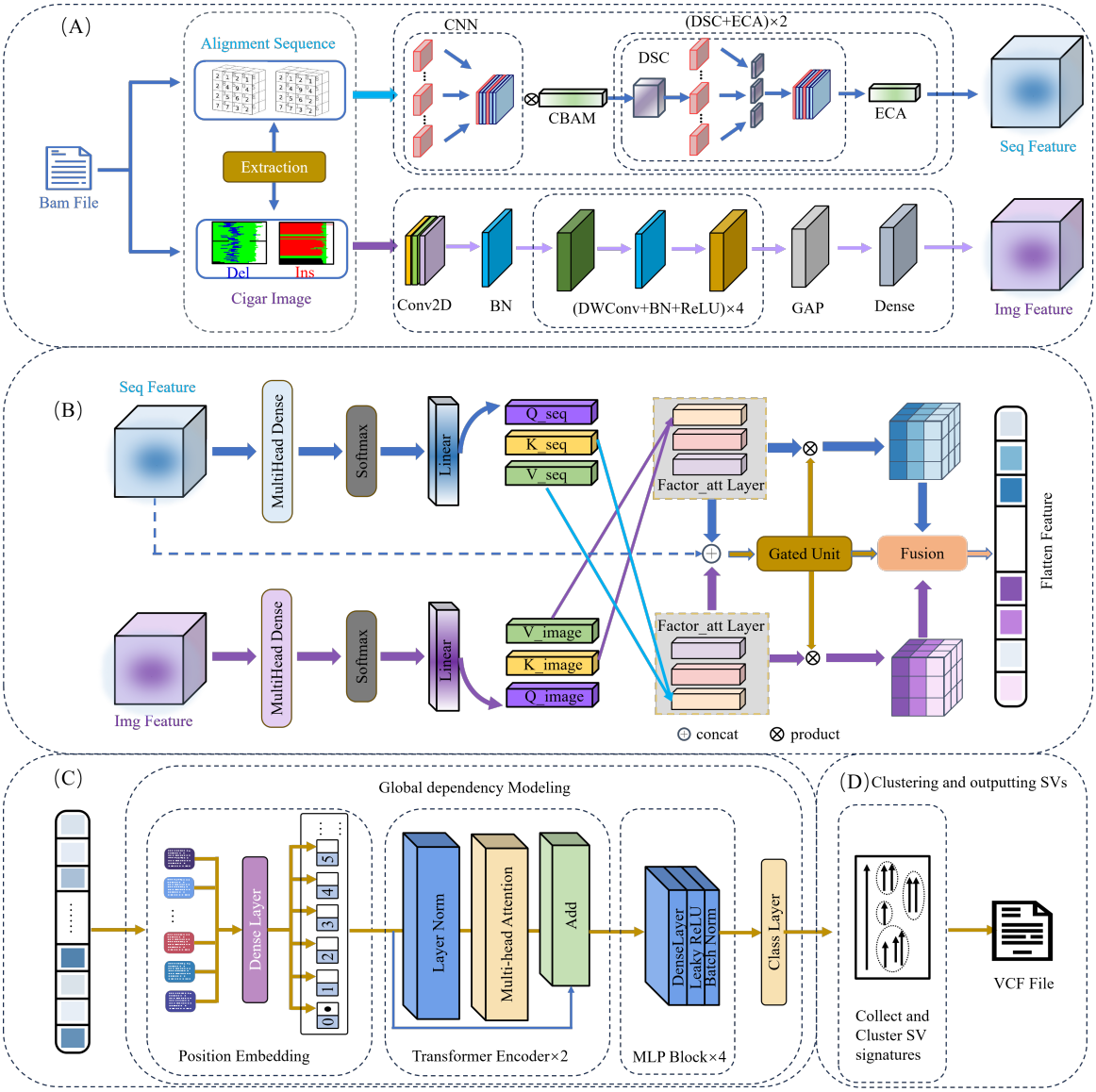

Fig. 1. Overall architecture of the TDSV framework. The pipeline comprises four core components: **(A) Dual-modal feature extraction** processes candidate SV regions by transforming sequence alignments into numerical feature matrices and converting CIGAR operations into image representations; **(B) Cross-modal fusion** employs bidirectional attention mechanisms to dynamically integrate features between sequence and image modalities; **(C) Global context modeling** utilizes a Transformer encoder to capture long-range dependencies across genomic windows for detecting complex structural variations; and **(D) Clustering and SV output** collects and clusters SV signatures, followed by filtering to generate high-confidence variant calls in VCF format.

C. Bidirectional Cross-Modal Attention Mechanism

While bidirectional attention paradigms have demonstrated efficacy in general computer vision tasks, their application to genomic variant calling remains unexplored. A critical challenge in this domain is distinguishing true variants from sequencing artifacts, which often requires synthesizing conflicting evidence from numerical alignment statistics and visual read pileups. To address this, TDSV introduces a tailored Bidirectional Cross-Modal Attention module. This module is not merely a feature aggregator but serves as a semantic bridge, enabling rigorous alignment between the abstract numerical tensor ($\mathbf{F}_{\text{seq}} \in \mathbb{R}^{d_{\text{seq}}}$) and the spatial CIGAR patterns ($\mathbf{F}_{\text{img}} \in \mathbb{R}^{d_{\text{img}}}$).

To operationalize this interaction, we project both modalities into a compatible latent space. Let $\mathcal{P}(\mathbf{X}, \mathbf{W}) = \mathbf{XW}$ denote the linear projection. We derive modality-specific queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) via learnable weight matrices $\mathbf{W}^{\{q,k,v\}}$. Specifically, the sequence-to-image pathway utilizes sequence features as queries ($\mathbf{Q}_{\text{seq}}$) to interrogate the image context ($\mathbf{K}_{\text{img}}, \mathbf{V}_{\text{img}}$), effectively "searching" for visual confirmation of statistical anomalies.

We define a unified Cross-Attention function $\Phi(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ to formulate this mechanism:

$$\Phi(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

where d_k is the scaling factor. By adapting this mechanism to

genomic data, we compute bidirectional contexts that capture complementary biological signals:

$$\mathbf{C}_{\text{seq}\to\text{img}} = \Phi(\mathbf{F}_{\text{seq}}\mathbf{W}^q_{\text{seq}}, \mathbf{F}_{\text{img}}\mathbf{W}^k_{\text{img}}, \mathbf{F}_{\text{img}}\mathbf{W}^v_{\text{img}}) \quad (2)$$

$$\mathbf{C}_{\text{img}\to\text{seq}} = \Phi(\mathbf{F}_{\text{img}}\mathbf{W}^q_{\text{img}}, \mathbf{F}_{\text{seq}}\mathbf{W}^k_{\text{seq}}, \mathbf{F}_{\text{seq}}\mathbf{W}^v_{\text{seq}}) \quad (3)$$

These pathways function synergistically: $\mathbf{C}_{\text{seq}\to\text{img}}$ leverages sequence semantics to identify discriminative regions within the noisy pileup images, while $\mathbf{C}_{\text{img}\to\text{seq}}$ uses visual patterns to filter false-positive signals in the sequence statistics. Recognizing that the reliability of each modality fluctuates across different genomic contexts (e.g., in regions with high alignment error rates), we implement an Adaptive Gating Fusion Unit. This unit constructs a holistic context vector $\mathbf{u}$ by concatenating the original sequence embedding with both cross-modal outputs:

$$\mathbf{u} = [\mathbf{F}_{\text{seq}} \parallel \mathbf{C}_{\text{seq}\to\text{img}} \parallel \mathbf{C}_{\text{img}\to\text{seq}}] \quad (4)$$

A learnable gating network then acts as a dynamic arbiter, generating sample-specific weights to modulate the fusion proportion based on the input features:

$$\mathbf{g} = \sigma(\mathbf{W}_g\mathbf{u} + \mathbf{b}_g) = [g_{\text{img}}, g_{\text{seq}}] \quad (5)$$

where σ denotes the Softmax function. The final fused representation $\bar{\mathbf{a}}$ is computed via weighted aggregation, ensuring that the model prioritizes the most informative modality for each specific variant candidate:

$$\bar{\mathbf{a}} = g_{\text{img}} \cdot \mathbf{C}_{\text{seq}\to\text{img}} + g_{\text{seq}} \cdot \mathbf{C}_{\text{img}\to\text{seq}} \quad (6)$$

Finally, to preserve the raw variant signals essential for precise genotyping and to prevent the "vanishing gradient" of original statistical information, we employ a residual connection. This step merges the sophisticated fused features with the original sequence input to form the comprehensive input $\mathbf{z}$ for the subsequent global transformer modeling:

$$\mathbf{z} = [\mathbf{F}_{\text{seq}} \parallel \bar{\mathbf{a}}] \quad (7)$$

D. Global Dependency Modeling with Transformer

To effectively capture long-range dependencies inherent in large insertion variants, we construct a sequence input $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{10}] \in \mathbb{R}^{10\times D}$ by assembling consecutive fused feature vectors. Each vector $\mathbf{z}_t$ corresponds to a 200-bp window, collectively spanning a 2-kb genomic region. To preserve the sequential order of these windows, a learnable positional encoding matrix $\mathbf{P} \in \mathbb{R}^{10\times D}$ is added to the input embeddings:

$$\mathbf{H}^{(0)} = \mathbf{Z} + \mathbf{P}. \quad (8)$$

The sequence is then processed by a stack of two Transformer encoder layers. Within each layer l, the input undergoes Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN), with residual connections and Layer Normalization applied as follows:

$$\mathbf{H}^{(l)\prime} = \text{LayerNorm}\left(\mathbf{H}^{(l-1)} + \text{MHSA}\left(\mathbf{H}^{(l-1)}\right)\right), \quad (9)$$

$$\mathbf{H}^{(l)} = \text{LayerNorm}\left(\mathbf{H}^{(l)\prime} + \text{FFN}\left(\mathbf{H}^{(l)\prime}\right)\right). \quad (10)$$

Specifically, the MHSA mechanism employs $H = 4$ attention heads, enabling each window to attend to all other windows in the sequence simultaneously, thereby capturing interactions across extended genomic distances. The FFN consists of two linear transformations with a non-linear activation, projecting the input to a hidden dimension of $4D$ to enhance representation capacity before projecting it back to D.

The final output of the Transformer encoder, $\mathbf{H}^{(L)} \in \mathbb{R}^{10\times D}$, serves as the input to the prediction head. It is passed through a 4-layer fully connected network (MLP) to extract high-level semantic features. Finally, a linear classifier followed by a sigmoid activation function predicts the SV probability for each window:

$$\hat{y}_t = \sigma(\mathbf{W}_{\text{cls}}\mathbf{h}^{(L)}_t + b_{\text{cls}}), \quad t = 1, \ldots, 10, \quad (11)$$

where $\mathbf{h}^{(L)}_t$ denotes the feature vector for the t-th window at the final layer L, $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{1\times D}$ is the weight matrix, and b_{cls} is the bias term. In the post-processing stage, adjacent windows with probabilities exceeding a threshold τ (set to 0.5) are merged to define the final candidate SV regions.

E. Estimating the insertion site and length

Following candidate sub-region identification, TDSV precisely estimates insertion breakpoints and lengths by integrating direct alignment signatures (CIGAR) and split-read alignments. First, insertion signatures with lengths ≥ 50 bp are extracted from CIGAR strings as triples $\mathcal{T} = (Chr, P_{\text{start}}, L_{\text{sv}})$. Adjacent triples on the same chromosome (within 30 bp) are merged to mitigate fragmentation caused by errors. Concurrently, soft-clipped reads are analyzed for split alignments; if two segments share the same chromosome and orientation, the junction boundaries are treated as candidate breakpoints. Finally, triples within 1500 bp are clustered. The final insertion site and length are determined by the median P_{start} and L_{sv} of each cluster, respectively, and exported in VCF format.

III. Results

To comprehensively evaluate our structural variant (SV) detection method on real biological data, we utilized two well-characterized human reference samples: HG002 (Genome in a Bottle) and NA19240. We processed long-read sequencing data from two platforms: PacBio Continuous Long Read (CLR) for both samples, and Circular Consensus Sequencing (CCS) exclusively for HG002. To strictly prevent data leakage, our models were trained on HG002 chromosomes 1–10, validated on chromosome 11, and tested on chromosomes 12–22. We systematically benchmarked our approach against three widely adopted, state-of-the-art SV callers: Sniffles2, SVIM, and cuteSV.

Performance was quantified using Truvari to calculate precision, recall, and F1-score. Furthermore, to assess model robustness across varying sequencing depths, we randomly downsampled the original high-coverage datasets. Specifically, the $69\times$ CLR dataset was subsampled to depths of $35\times$, $20\times$, $10\times$, and $5\times$, while the $28\times$ CCS data was reduced to $10\times$ and $5\times$ subsets.

A. Benchmark Results on CLR Data Across Coverages

We conducted comprehensive benchmark evaluations comparing TDSV against Sniffles2, SVIM, and cuteSV using the HG002 sample. As illustrated in Fig. 2, TDSV demonstrated superior performance across all tested coverage levels. Under the original high-coverage setting (69× PacBio CLR), TDSV achieved an F1-score of 92.82% for insertion detection (Fig. 2B), significantly outperforming Sniffles2 (76.43%), SVIM (86.25%), and cuteSV (90.97%). This result establishes TDSV's strong baseline capability in ideal sequencing conditions.

The robustness of TDSV became increasingly evident in downsampled datasets. At 35× coverage (Fig. 2C), TDSV attained the highest F1-score of 89.26% alongside the highest recall among all methods. As coverage was reduced to 20× (Fig. 2D), TDSV surpassed the second-best tool, cuteSV, by 1.33% in F1-score while securing the top precision. This trend continued at 10× coverage (Fig. 2E), where TDSV outperformed cuteSV by a margin of 2.85% in F1-score.

Detecting SVs becomes particularly challenging in low-coverage scenarios due to data sparsity and the higher signal-to-noise ratio required to distinguish true variants from sequencing errors. Notably, even at the extremely low coverage of 5× (Fig. 2F), TDSV maintained its lead over all competing tools. This stability is largely attributed to the strong representational capacity of its Transformer-based architecture, which effectively captures structural patterns even with limited read support.

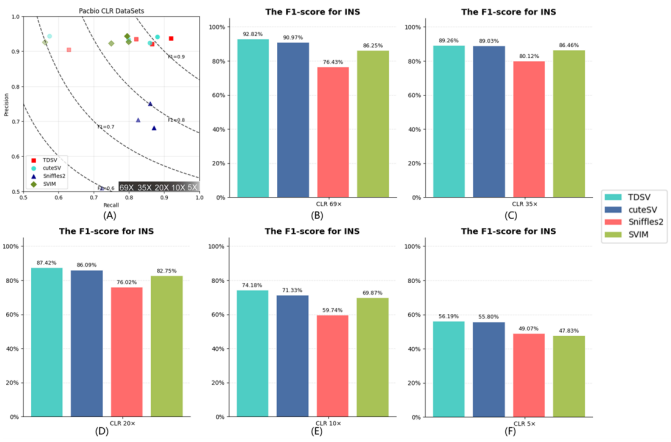

Fig. 2. Performance comparison on PacBio CLR dataset. (A) Performance of each detection tool across different coverage levels. (B) F1-scores for insertion detection at 69× coverage. (C) F1-scores at 35× coverage. (D) F1-scores at 20× coverage. (E) F1-scores at 10× coverage. (F) F1-scores at 5× coverage.

B. Benchmark Results on CCS Data Across Coverages

To further evaluate the robustness of our method, we benchmarked TDSV against Sniffles2, SVIM, and cuteSV using PacBio CCS data downsampled to 10× and 5× coverage. TDSV consistently achieved the best performance across all three coverage levels. Specifically, on the 28× dataset, TDSV maintained superior performance with an F1-score of 93.22%. Notably, at 10× coverage, its F1-score surpassed that of Sniffles2 by 1.66%, and at 5× coverage, it outperformed the second-best method by nearly 4%. These results demonstrate that TDSV delivers consistently robust performance on CCS data, even under low-coverage conditions.

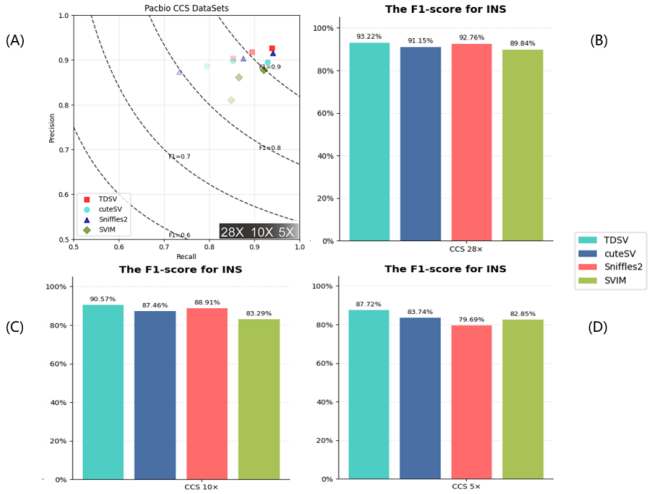

Fig. 3. Performance comparison on PacBio CCS dataset. (A) Performance of each detection tool across different coverage levels.(B–D) F1-score for (B) 28×, (C) 10×, and (D) 5× coverage datasets.

C. Performance on NA19240 Dataset

To assess the generalizability of TDSV beyond the standard HG002 genome, we benchmarked our method on the NA19240 dataset (PacBio CLR; ∼6,503 bp mean read length, ∼41× coverage) utilizing insertion ground truths from NCBI dbVar. As presented in Table I, TDSV delivered the most competitive performance among the evaluated tools, attaining the highest F1-score of 31.55%—an absolute improvement of 4.13% over the second-best method, cuteSV. Furthermore, TDSV yielded the highest precision (68.26%) and recall (20.51%) across all baseline comparisons. These findings confirm the consistent robustness and adaptability of TDSV when applied to diverse biological samples and complex genomic contexts.

TABLE I
PERFORMANCE COMPARISON ON THE NA19240 DATASET

Metric	TDSV	cuteSV	Sniffles2	SVIM
Precision	**0.6826**	0.6276	0.4915	0.4163
Recall	**0.2051**	0.1754	0.1822	0.0305
F1-score	**0.3155**	0.2742	0.2659	0.0568

D. Ablation Study on Multimodal Architecture

To evaluate the contribution of TDSV's multimodal architecture, we conducted an ablation study comparing the full model (`SeqAndPic`) against a sequence-statistics-only baseline (`Seq`). Both configurations underwent a consistent 20-epoch training schedule with early stopping to ensure fair

comparison. As shown in Fig. 4A, the multimodal model (red) consistently outperforms the unimodal baseline (green) in TP ratios across both training and validation sets, indicating that integrating visual features significantly enhances sensitivity. Furthermore, the loss landscape (Fig. 4B) reveals that `SeqAndPic` converges more rapidly and stabilizes at lower loss levels, whereas the `Seq` model exhibits higher loss and instability in later epochs. These results demonstrate that multimodal integration not only improves representation capability but also enhances generalization and training stability.

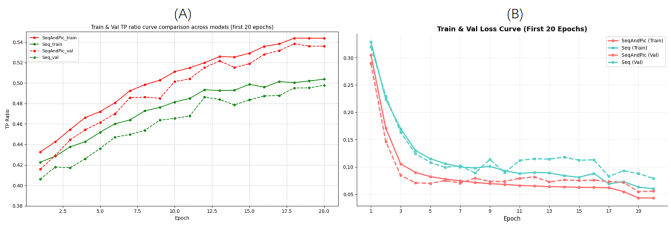

Fig. 4. Ablation study comparing the performance of multimodal and sequence-only approaches on PacBio CLR 69× data. (A) Comparison of training and validation True Positive (TP) ratio curves. The multimodal model (`SeqAndPic`) achieves consistently higher TP ratios. (B) Comparison of training and validation loss curves, highlighting the faster convergence and stability of the multimodal approach.

IV. Conclusion

In this work, we proposed TDSV, a novel Transformer-based dual-modal framework designed to address the limitations of existing single-modality approaches in structural variant detection. Our method features three key innovations: (1) a comprehensive dual-modal representation integrating sequence-level statistical properties with CIGAR-derived visual patterns; (2) a bidirectional cross-modal attention mechanism facilitating dynamic feature alignment; and (3) a Transformer-based global dependency module to effectively capture long-range genomic contexts.

Extensive benchmarking on HG002 samples across PacBio CLR and CCS platforms demonstrates our approach's superiority. TDSV consistently outperformed state-of-the-art methods, achieving higher F1-scores and recall rates. Notably, it exhibited remarkable robustness in low-coverage scenarios (e.g., $5\times$), where traditional heuristic methods often struggle due to data sparsity and sequencing noise.

The success of TDSV underscores the efficacy of leveraging complementary multi-modal information. By employing adaptive gating to weight modality contributions, our framework effectively mitigates alignment artifacts and enhances detection precision. In future work, we plan to extend TDSV to broader SV types (e.g., deletions and inversions) and incorporate population-scale genotypes. Furthermore, future evaluations will include direct comparisons with emerging deep learning-based variant callers to further benchmark our multimodal architecture.

References

[1] M. Kolmogorov, K. J. Billingsley, M. Mastoras, et al., "Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation," *Nat. Methods*, vol. 20, no. 10, pp. 1483-1492, Oct. 2023.

[2] S. S. Ho, A. E. Urban, and R. E. Mills, "Structural variation in the sequencing era," *Nat. Rev. Genet.*, vol. 21, no. 3, pp. 171-189, Mar. 2020.

[3] S. Pande, M. Dawood, and C. M. Grochowski, "Structural Variants: Mechanisms, Mapping, and Interpretation in Human Genetics," *Genes*, vol. 16, no. 8, p. 905, Aug. 2025, doi: 10.3390/genes16080905.

[4] M. E. MacDonald, C. M. Ambrose, M. P. Duyao, et al., "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes," *Cell*, vol. 72, no. 6, pp. 971-983, Mar. 1993.

[5] S. A. Bustin and K. A. Jellinger, "Advances in Molecular Medicine: Unravelling Disease Complexity and Pioneering Precision Healthcare," *Int. J. Mol. Sci.*, vol. 24, no. 18, p. 14168, Sep. 2023, doi: 10.3390/ijms241814168.

[6] C. Scarano, I. Veneruso, R. R. De Simone, G. Di Bonito, A. Secondino, and V. D'Argenio, "The third-generation sequencing challenge: Novel insights for the omic sciences," *Biomolecules*, vol. 14, no. 5, p. 568, May 2024.

[7] K. Athanasopoulou, M. A. Boti, P. G. Adamopoulos, P. C. Skourou, and A. Scorilas, "Third-generation sequencing: The spearhead towards the radical transformation of modern genomics," *Life*, vol. 12, no. 1, p. 30, Jan. 2022.

[8] B. Lu, Z. Guo, X. Liu, et al., "Comprehensive comparison of the third-generation sequencing tools for bacterial 6mA profiling," *Nat. Commun.*, vol. 16, p. 3982, Jun. 2025.

[9] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, "Long-read human genome sequencing and its applications," *Nat. Rev. Genet.*, vol. 21, no. 10, pp. 597-614, Oct. 2020.

[10] Y.-H. Liu, C. Luo, S. G. Golding, et al., "Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data," *Nat. Commun.*, vol. 15, no. 1, p. 2447, Mar. 2024.

[11] T. Jiang, Y. Liu, Y. Jiang, et al., "Long-read-based human genomic structural variation detection with cuteSV," *Genome Biol.*, vol. 21, no. 1, p. 189, Jun. 2020.

[12] D. Heller and M. Vingron, "SVIM: structural variant identification using mapped long reads," *Bioinformatics*, vol. 35, no. 17, pp. 2907-2915, Sep. 2019.

[13] Y. Chen, A. Y. Wang, C. A. Barkley, et al., "Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak," *Nat. Commun.*, vol. 14, p. 283, Jan. 2023.

[14] C. Liu, F. Xu, C. Gao, Z. Wang, Y. Li, and J. Gao, "Deep learning resilience inference for complex networked systems," *Nat. Commun.*, vol. 15, p. 9203, Oct. 2024, doi: 10.1038/s41467-024-53303-4.

[15] Y.-Y. Liu, K. Cheng, R. Just, S. Enke, and J.-A. Bright, "Sequencing-induced artefacts in NGS STR data," *Forensic Sci. Int., Genet.*, vol. 72, p. 103086, 2024, doi: 10.1016/j.fsigen.2024.103086.

[16] Z. Liu, Z. Xie, and M. Li, "Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data," *Genome Biol.*, vol. 25, no. 1, p. 188, Mar. 2024.

[17] A. A. Helal, B. T. Saad, M. T. Saad, et al., "Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data," *Sci. Rep.*, vol. 14, p. 6160, Mar. 2024.

[18] Y. Zheng and X. Shang, "SVcnn: an accurate deep learning-based method for detecting structural variation based on long-read data," *BMC Bioinf.*, vol. 24, no. 1, p. 213, Aug. 2023.

[19] R. Gao, J. Luo, H. Ding, et al., "INSnet: a method for detecting insertions based on deep learning network," *BMC Bioinf.*, vol. 24, no. 1, p. 80, Feb. 2023.

[20] H. Wang, C. Li, X. Yu, and J. Gao, "Deletion variants calling in third-generation sequencing data based on a dual-attention mechanism," *Brief. Bioinform.*, vol. 25, no. 4, Jul. 2024.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3-19.

[22] Q. Wang, B. Wu, P. Zhu, et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11534-11542.