SIGNRET: SIGN LANGUAGE RETRIEVAL VIA FINE-GRAINED GESTURE DESCRIPTIONS AND RANKING REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign language retrieval aims to establish bidirectional correspondence between sign language videos and natural language sentences. Existing methods typically match global video–text representations or aggregate local clip–word similarities. Global matching may overlook subtle gesture differences, while local matching relies on how well these details are encoded in the visual features. Moreover, both paradigms typically determine the final ranking from independently computed query–candidate similarities, without explicitly modeling the score and rank structure of the retrieved candidate list. Consequently, they may struggle to identify the correct match when multiple candidates exhibit similar visual patterns. To address these limitations, we propose SignRet, a sign language retrieval framework consisting of Sign Language Description Enhancement and Candidate Ranking Refinement. The former incorporates frame-level gesture descriptions into temporally corresponding visual features, enriching video representations with explicit fine-grained cues for more discriminative cross-modal matching. The latter revisits the top candidates obtained from the initial matching and models their contextual relationships to recalibrate candidate relevance, thereby improving the reliability of the final retrieval results. Experiments on PHOENIX-2014T, How2Sign, and CSL-Daily demonstrate that SignRet consistently improves both retrieval directions and achieves strong performance.

1 INTRODUCTION

Sign language is a natural visual language that conveys meaning through coordinated hand gestures, body movements, and facial expressions. Recent advances have substantially promoted research on sign language recognition, translation, and production (Jang et al., 2025; Guan et al., 2025; Guo et al., 2025; Li et al., 2025b; Liu et al., 2026). However, most existing studies focus on understanding or generating sign language content within specific datasets, while the organization and efficient access of large-scale sign language video collections remain underexplored. The rapidly growing volume of sign language content on educational, social, and media platforms makes such capabilities increasingly important. This motivates sign language retrieval (Cheng et al., 2023; Duarte et al., 2022; Jiang et al., 2024a; Lee et al., 2026), a cross-modal task that establishes bidirectional correspondence between sign language videos and natural language sentences. Specifically, text-to-sign-language-video retrieval searches a video collection for the sign language video corresponding to a textual query, whereas sign-language-video-to-text retrieval identifies the corresponding natural language sentence for a given sign language video.

Research on sign language retrieval remains limited. A closely related but distinct task is generic text–video retrieval (Zhao et al., 2026; Shen et al., 2025; Cho et al., 2025; Bai et al., 2025), which retrieves semantically relevant videos from textual queries or vice versa. Prevailing approaches fall into two categories: global representation matching and fine-grained clip–token similarity aggregation. However, sign language retrieval requires particular sensitivity to fine-grained gesture details. While objects, scenes, and coarse actions provide important cues in generic video retrieval, sign language meaning can depend on subtle differences in handshape, hand location, body posture, and facial expression. These differences require visual representations that preserve discriminative gesture cues.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

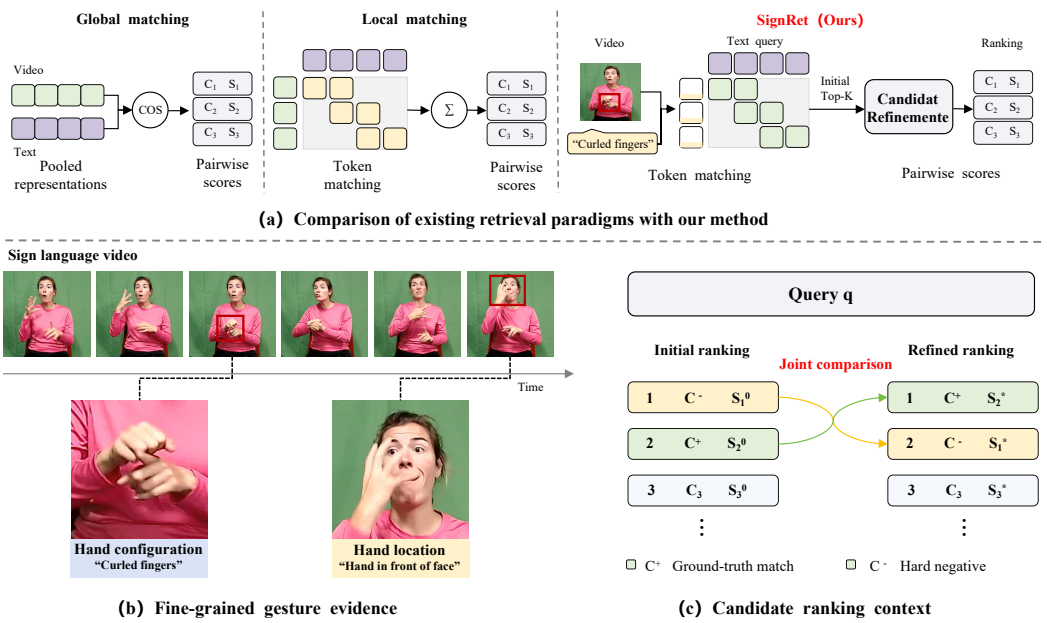


Figure 1: Motivation of our proposed SignRet. (a) Comparison of three retrieval paradigms. (b) Frame-level descriptions provide fine-grained gesture cues. (c) Listwise score refinement adjusts candidate ordering.

Existing sign language retrieval methods largely follow these two paradigms, as illustrated in Figure 1(a). Global matching methods encode an entire sign language video and sentence into compact representations and rank candidates by their global similarities (Duarte et al., 2022; Jiang et al., 2024b; Chen et al., 2025). Local matching methods instead retain video clips and text tokens, estimate their fine-grained correspondences, and aggregate the local evidence into a retrieval score (Cheng et al., 2023; Jiang et al., 2024a). Despite their progress, these methods have two limitations. First, global representations may obscure subtle gesture differences, while local matching depends on how well these differences are encoded in the visual features. Frame-level descriptions can make attributes such as handshape and hand location explicit, providing complementary cues for matching, as illustrated in Figure 1(b). Second, both paradigms typically rank candidates using independently computed query–candidate similarities, without explicitly modeling the score and rank structure of the retrieved list. When several candidates receive similar scores, their relative positions and score gaps provide additional context for ranking refinement, as illustrated in Figure 1(c). These limitations motivate us to enhance fine-grained gesture representations and use candidate-list context to refine retrieval scores.

To address these challenges, we propose SignRet, a sign language retrieval framework that improves both fine-grained gesture representation and candidate ranking. First, Sign Language Description Enhancement employs a frozen multimodal large language model (MLLM) (Yang et al., 2025; Zhang et al., 2025) to interpret individual video frames and generate frame-level gesture descriptions. The textual query and these descriptions provide complementary semantics at different granularities: the former summarizes the sentence-level meaning of the video, while the latter captures frame-level hand configurations, spatial locations, and body and facial states. For each RGB clip, we collect the descriptions of its temporally corresponding frames and integrate the relevant gesture cues through local visual–description interaction. This produces gesture-enhanced clip representations for fine-grained word–clip matching. Second, Candidate Ranking Refinement (CRR) constructs matching features for the initially retrieved top candidates using relative scores and ranks, reverse matching statistics, and global similarities among text, video, and description representations. A listwise encoder jointly processes these features and predicts residual score corrections, allowing each candidate's adjustment to depend on the matching evidence of its competitors. Separate models predict corrections for T2V and V2T, which are combined into the final similarity matrix. Our main contributions are summarized as follows:

- We propose SignRet, a sign language retrieval framework that jointly improves fine-grained gesture representation and candidate-aware ranking for reliable retrieval.
- We introduce Sign Language Description Enhancement, which uses frame-level gesture descriptions to capture fine-grained hand, body, and facial cues and integrates them into temporally corresponding RGB clips through local visual–description interaction.
- We introduce Candidate Ranking Refinement, which jointly processes candidate matching features to predict residual corrections to the initial retrieval scores.
- Extensive experiments on three public datasets demonstrate that SignRet consistently improves both retrieval directions and achieves strong performance.

2 Related Work

2.1 Text–Video Retrieval

Text–video retrieval ranks videos and natural language descriptions by learning cross-modal representations and matching functions. CLIP4Clip (Luo et al., 2022) adapted image–text pretraining to videos by aggregating frame-level CLIP representations. UCoFiA (Wang et al., 2023) further unified coarse-to-fine alignments across video–sentence, clip–phrase, and frame–word levels. More recently, Video-ColBERT (Reddy et al., 2025) introduced contextualized late interaction to preserve token-level video–text correspondences for efficient retrieval. MonSTeR (Collorone et al., 2025) jointly modeled motion, scene, and text information in a unified framework to support retrieval across different visual semantics. Auxiliary video descriptions have also been explored to provide explicit semantic cues. Cap4Video (Wu et al., 2023) generated video-level captions and exploited them for data augmentation, video–caption interaction, and complementary similarity estimation. Ventura et al. (Ventura et al., 2025) used image captioners to annotate individual video frames and employed the generated descriptions as supervision for retrieval learning. NarVid (Hur et al., 2025) generated a sequence of frame-level captions as a video narration and incorporated them through feature enhancement, query-aware filtering, and dual-modal matching. CoreCaption (Jang, 2026) further extracted representative captions from multiple generated descriptions to emphasize essential video content. Cross-modal retrieval can also be improved by reranking the results of an initial matching stage. Wang et al. (Wang et al., 2019) used multimodal tensor fusion to capture rich image–text interactions and rerank the initially retrieved results, while Yuan et al. (Yuan et al., 2022) combined global and local information for remote-sensing retrieval and subsequent reranking.

2.2 Sign Language Retrieval

Sign language retrieval aims to establish bidirectional semantic correspondence between sign language videos and natural language sentences. Duarte et al. (Duarte et al., 2022) introduced free-form text-based sign language video retrieval and proposed SPOT-ALIGN, which alternates sign spotting and feature alignment to expand the available training supervision. CiCo (Cheng et al., 2023) formulated the task from a cross-lingual perspective and learned fine-grained correspondences between sign language videos and words through domain-aware contrastive learning. SEDS (Jiang et al., 2024a) combined RGB and pose representations to capture complementary global semantics and local motion details, together with context-aware fusion and fine-grained matching. C^2RL (Chen et al., 2025) jointly modeled signing content and linguistic context to improve gloss-free representations for both retrieval and translation. SignCLIP (Jiang et al., 2024b) further explored large-scale multilingual sign-language–text contrastive pretraining to obtain transferable representations for retrieval and other sign language tasks. Lee et al. (Lee et al., 2026) investigated the mismatch between semantic and visual hardness and introduced sign-aware hard-negative mining to improve candidate discrimination. SignRet complements these methods with temporally aligned gesture-description fusion and listwise score refinement.

3 Methodology

This section presents the SignRet framework. We first formulate the sign language retrieval task and define the notation. We then introduce Sign Language Description Enhancement (SDE), which enriches RGB clip features with temporally aligned frame-level gesture descriptions. Next, we present

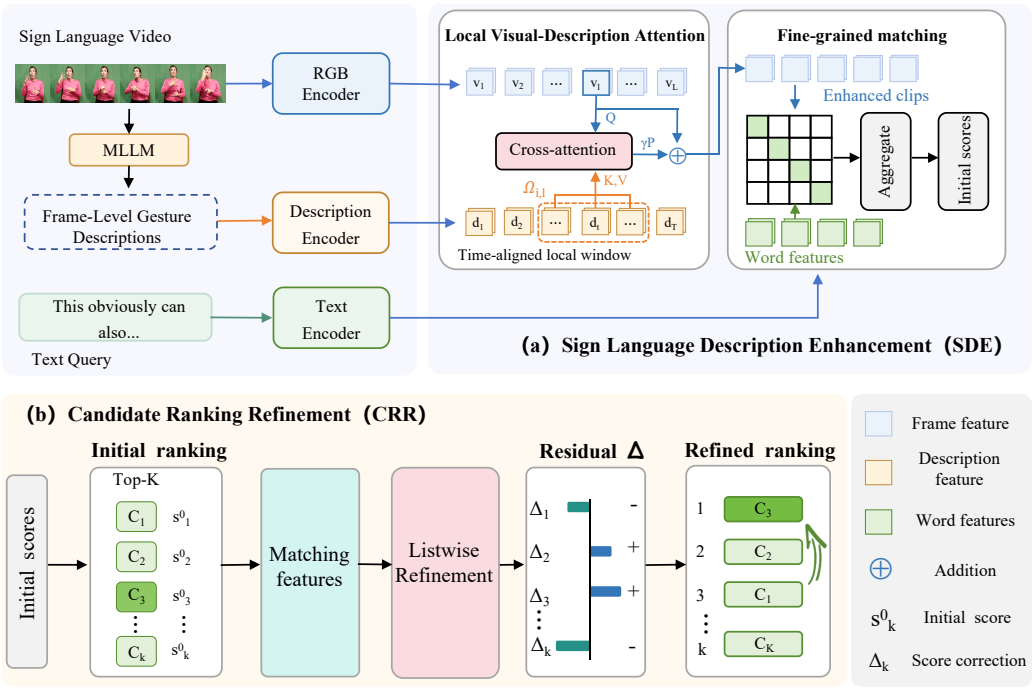


Figure 2: Overall framework of our proposed SignRet.

Candidate Ranking Refinement (CRR), which jointly encodes the matching features of the top-ranked candidates to predict residual score corrections. Finally, we describe the training objectives and inference procedure. The overall framework is illustrated in Figure 2.

3.1 Problem Formulation

We consider retrieval between sign language videos and natural language sentences. In text-to-video (T2V) retrieval, a sentence is used to rank candidate videos. In video-to-text (V2T) retrieval, a video is used to rank candidate sentences. For a video–text pair, we represent the video as a sequence of RGB clip features $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_L]$ and the sentence as a sequence of token features $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$. Here, L and M are the numbers of video clips and text tokens, respectively. We omit sample indices when describing the operations on an individual video–text pair. After description enhancement and video–text matching, the initial scores form a matrix $\mathbf{S}^0 \in \mathbb{R}^{N_v \times N_t}$, where N_v and N_t are the numbers of videos and sentences. Each entry S^0_{ij} measures the similarity between video i and sentence j. A column contains the candidate scores for a T2V query, and a row contains those for a V2T query. CRR updates this matrix, and the resulting scores are used to rank candidates in both directions.

3.2 Sign Language Description Enhancement

Sign language videos with similar overall appearances may differ in handshape, hand location, or body posture. These details are important for video–text matching but may be underrepresented in RGB features. We introduce SDE to supplement RGB clip features with frame-level gesture descriptions from the corresponding temporal regions.

Frame-level descriptions. Following MMSLT (Kim et al., 2025), we use a frozen MLLM with the same prompt and generation settings to describe each sampled frame independently. Neither paired retrieval sentences nor gloss annotations are provided. The descriptions characterize visible gesture states. They are generated offline and encoded with BERT into a cached sequence $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_T]$, where T is the number of described frames.

Local fusion and matching. Each RGB clip spans a short temporal segment and may correspond to several frame-level descriptions. We use the frame timestamps to select descriptions within the clip's local temporal window, whose indices form Ω_l. This keeps the description context local to the visual content being enhanced. Within this window, the clip feature acts as the query, and the description features act as keys and values. Cross-attention weights the descriptions according to their compatibility with the clip feature and combines them into a description context

$$\mathbf{c}_l = \mathrm{CrossAttn}\left(\mathbf{v}_l,\, \mathbf{D}_{\Omega_l},\, \mathbf{D}_{\Omega_l}\right), \tag{1}$$

where $\mathbf{D}_{\Omega_l}$ contains the description features selected for clip l, and the three arguments denote the query, keys, and values, respectively.

We project this context back to the RGB feature dimension and add it to the original clip feature

$$\widetilde{\mathbf{v}}_l = \mathbf{v}_l + \gamma \mathcal{P}(\mathbf{c}_l), \tag{2}$$

where $\mathcal{P}$ is a learned projection and γ controls the contribution of the description context. Repeating this operation for all clips gives the enhanced video features $\widetilde{\mathbf{V}} = [\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_L]$.

For video–text matching, we compare each text token with each video clip using cosine similarity

$$A_{ml} = \frac{\mathbf{w}_m^\top \widetilde{\mathbf{v}}_l}{\|\mathbf{w}_m\|_2 \|\widetilde{\mathbf{v}}_l\|_2}. \tag{3}$$

The resulting matrix $\mathbf{A} \in \mathbb{R}^{M \times L}$ contains the similarities between all M text tokens and L video clips, with rows corresponding to text tokens and columns to video clips. We aggregate these local similarities using the fine-grained matching rule of SEDS (Jiang et al., 2024a) to obtain an initial video–text score. Applying this procedure to all pairs yields $\mathbf{S}^0$, where S^0_{ij} is the initial score between video i and sentence j.

3.3 Candidate Ranking Refinement

Several candidates may receive similar scores even after description-enhanced matching. CRR combines relative scores, reverse matching statistics, and global similarities to refine their ordering.

Matching features. For a text query, let $c_1, \ldots, c_K$ be the initial Top-K video candidates, with scores $s^0_1, \ldots, s^0_K$ in descending order. We first characterize each match relative to the score distributions in both retrieval directions

$$z_k = \frac{s^0_k - \mu}{\sigma}, \qquad z'_k = \frac{s^0_k - \mu_k}{\sigma_k}. \tag{4}$$

Here, μ and σ are the mean and standard deviation of the query's scores against all videos, while μ_k and σ_k are those of candidate video c_k's scores against all texts. Statistics use the complete retrieval set of the current split, with standard deviations floored at 10^{-6}. The two scores express the match's strength relative to competing videos and competing texts, respectively.

To further describe the candidate's position in the ranking, we include its full-gallery rank, its gap to the leading candidate $s^0_1 - s^0_k$, and the shared gap between the two leading candidates $s^0_1 - s^0_2$. The leading scores are taken from the selected list.

Besides, we incorporate similarities between global content representations

$$\mathbf{g}_k = \left[\cos(\bar{\mathbf{w}}, \bar{\mathbf{v}}_k),\, \cos(\bar{\mathbf{w}}, \bar{\mathbf{d}}_k),\, \cos(\bar{\mathbf{v}}_k, \bar{\mathbf{d}}_k)\right], \tag{5}$$

where $\bar{\mathbf{w}}$, $\bar{\mathbf{v}}_k$, and $\bar{\mathbf{d}}_k$ are the global representations of the query sentence, enhanced video, and video descriptions. We concatenate the relative matching features, reverse matching statistics, and global similarities into $\mathbf{x}_k$. The same construction applies to video queries.

Listwise refinement. We project the matching vectors into embeddings and process them jointly with a self-attention encoder. Each output $\mathbf{h}_k$ incorporates information from the candidate's own features and those of its competitors. A shared prediction head g produces scalar adjustments, which are centered within the list

$$\Delta_k = g(\mathbf{h}_k) - \frac{1}{K} \sum_{r=1}^{K} g(\mathbf{h}_r). \tag{6}$$

Centering removes a common offset while preserving the differences between candidate adjustments. The T2V and V2T models have the same architecture but independent parameters. For a video–text pair (i, j), they predict $\Delta_{ij}^{\mathrm{T2V}}$ from the video candidate list of sentence j and $\Delta_{ij}^{\mathrm{V2T}}$ from the text candidate list of video i, respectively. A correction is zero if the pair is outside the selected list in that direction. The final score combines both corrections

$$S_{ij}^{*} = S_{ij}^{0} + \alpha \left(\Delta_{ij}^{\mathrm{T2V}} + \Delta_{ij}^{\mathrm{V2T}} \right), \tag{7}$$

where α controls the correction strength.

3.4 Training Objectives and Inference

Retrieval training. We first train the description-enhanced retrieval model to distinguish matched video–text pairs from mismatched pairs. For a batch of B paired examples, let S_{ij}^{0} denote the similarity between video i and sentence j. The retrieval objective is $\mathcal{L}_{\mathrm{ret}} = \frac{1}{2}(\mathcal{L}_{\mathrm{T2V}} + \mathcal{L}_{\mathrm{V2T}})$, where

$$\mathcal{L}_{\mathrm{T2V}} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\exp(S_{jj}^{0}/\tau)}{\sum_{i=1}^{B} \exp(S_{ij}^{0}/\tau)}, \qquad \mathcal{L}_{\mathrm{V2T}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{ii}^{0}/\tau)}{\sum_{j=1}^{B} \exp(S_{ij}^{0}/\tau)}. \tag{8}$$

Here, $\tau > 0$ is the temperature. The two losses encourage matched video–text pairs to score higher than mismatched pairs in both retrieval directions.

CRR training. We freeze the first-stage retrieval model, including SDE, and train CRR on its candidate lists. For a training list $\mathcal{C}$ containing the matched candidate, let $s_k = s_k^0 + \alpha\Delta_k$ denote the corrected score in the current retrieval direction. We write s_+ for the matched candidate's score and s_- for the highest corrected score among the negatives.

The listwise classification loss encourages the matched candidate to rank above the other candidates

$$\ell_{\mathrm{ce}} = -\log \frac{\exp(s_+)}{\sum_{k=1}^{|\mathcal{C}|} \exp(s_k)}. \tag{9}$$

We further constrain its score relative to the hardest negative using a margin loss and a smooth ranking loss

$$\ell_{\mathrm{mar}} = [\delta + s_- - s_+]_+, \qquad \ell_{\mathrm{rank}} = \tau_r \log\left(1 + \exp\left(\frac{s_- - s_+}{\tau_r}\right)\right). \tag{10}$$

Here, $[x]_+ = \max(0, x)$ and $\delta > 0$ is the required margin. The margin loss penalizes score gaps below δ, while the smooth ranking loss continues to decrease as the positive score moves above the hardest negative. The parameter $\tau_r > 0$ controls its smoothness.

The CRR objective is a weighted sum of the three losses

$$\mathcal{L}_{\mathrm{CRR}} = \lambda_{\mathrm{ce}}\mathcal{L}_{\mathrm{ce}} + \lambda_{\mathrm{mar}}\mathcal{L}_{\mathrm{mar}} + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}}, \tag{11}$$

where each loss is averaged over training queries. The weights λ_{ce}, λ_{mar}, and λ_{rank} control the contributions of listwise classification, margin separation, and smooth ranking, respectively.

Inference. The retrieval model first computes the initial similarity matrix using the cached frame descriptions. The T2V and V2T CRR models then process their respective Top-K lists and predict score corrections. Unlike the direction-specific scores used for CRR training, the final inference score combines corrections from both directions according to Eq. 7. We rank videos within each column of the resulting matrix for T2V retrieval and sentences within each row for V2T retrieval.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SignRet on three public sign language datasets. PHOENIX-2014T (Camgoz et al., 2018b) contains German Sign Language weather broadcasts paired with German sentences, with 7,096/519/642 training/validation/test pairs. How2Sign (Duarte et al., 2021) provides

Table 1: Comparison with state-of-the-art methods on PHOENIX-2014T, How2Sign, and CSL-Daily. Checkmarks indicate the visual modalities used by each method. The best and second-best results are highlighted in **bold** and underlined, respectively. mR@1 denotes the average of T2V R@1 and V2T R@1. Gray rows denote SignRet under different visual-input settings, and SEDS* denotes our reproduced results. "–" indicates unavailable or unverified values.

Method	Visual Input		Text-to-Video (T2V)					Video-to-Text (V2T)					mR@1↑
	RGB	Pose	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	
PHOENIX-2014T													
SA-SR	✓		30.2	53.1	63.4	4.5	–	28.8	52.0	60.8	–	–	29.5
SA-CM	✓		48.6	76.5	84.6	2.0	–	50.3	78.4	84.4	–	–	49.5
SA-COMB	✓		55.8	79.6	92.1	**1.0**	–	53.1	79.4	86.1	**1.0**	–	54.5
GFSLT-VLP	✓		67.9	88.4	93.8	**1.0**	–	69.4	88.7	93.3	**1.0**	–	68.7
CiCo	✓		69.5	86.6	92.1	**1.0**	–	70.2	88.0	92.8	**1.0**	–	69.9
UPRet	✓		72.0	89.1	94.1	**1.0**	4.4	72.0	89.4	93.3	**1.0**	4.6	72.0
SEDS*	✓	✓	75.4	91.6	**95.8**	**1.0**	**3.5**	76.3	**92.7**	**95.5**	**1.0**	3.1	75.9
SignRet	✓		73.7	88.0	91.9	**1.0**	4.5	73.7	88.3	92.4	**1.0**	4.4	73.7
SignRet	✓	✓	**79.9**	**92.2**	95.6	**1.0**	**3.5**	**80.1**	92.5	**95.5**	**1.0**	**3.0**	**80.0**
How2Sign													
SA-SR	✓		18.9	32.1	36.5	62.0	–	11.6	27.4	32.5	69.0	–	15.3
SA-CM	✓		24.3	40.7	46.5	16.0	–	17.9	40.1	46.9	14.0	–	21.1
SA-COMB	✓		34.2	48.0	52.6	8.0	–	23.6	47.0	53.0	7.5	–	28.9
CiCo	✓		56.6	69.9	74.7	**1.0**	–	51.6	64.8	70.1	**1.0**	–	54.1
UPRet	✓		59.1	71.5	75.7	**1.0**	54.4	53.4	65.4	70.0	**1.0**	76.4	56.3
SEDS*	✓	✓	61.7	75.1	79.6	**1.0**	44.9	57.1	69.8	**75.3**	**1.0**	**59.1**	59.4
SignRet	✓		60.3	71.2	76.0	**1.0**	53.8	54.1	65.7	69.5	**1.0**	72.5	57.2
SignRet	✓	✓	**63.7**	**76.0**	**80.3**	**1.0**	**44.7**	**59.0**	**70.5**	**75.3**	**1.0**	**59.1**	**61.4**
CSL-Daily													
CiCo	✓		75.3	88.2	91.9	**1.0**	–	74.7	89.4	92.2	**1.0**	–	75.0
UPRet	✓		78.4	89.1	92.0	**1.0**	6.7	77.0	89.2	92.7	**1.0**	5.5	77.7
SEDS*	✓	✓	84.5	94.1	95.5	**1.0**	3.3	84.1	94.0	95.8	**1.0**	3.4	84.3
SignRet	✓		80.8	90.2	93.0	**1.0**	6.4	77.5	88.8	92.6	**1.0**	5.2	79.2
SignRet	✓	✓	**88.7**	**94.6**	**95.8**	**1.0**	**3.0**	**87.0**	**94.2**	**96.1**	**1.0**	**3.3**	**87.8**

over 80 hours of American Sign Language recordings and 35,191 sentence-level clips with English annotations. CSL-Daily (Zhou et al., 2021) contains Chinese Sign Language videos paired with Chinese sentences covering daily-life topics such as shopping, travel, and medical care.

Evaluation Metrics. We evaluate T2V and V2T retrieval using R@1, R@5, R@10, Median Rank (MedR), Mean Rank (MnR), and mean R@1 (mR@1). R@K is the percentage of queries whose ground-truth match appears among the top K retrieved candidates, with higher values indicating better performance. MedR and MnR denote the median and mean ranks of the ground-truth matches, respectively, with lower values being better. mR@1 is the average of T2V R@1 and V2T R@1.

Implementation Details. RGB features are extracted using I3D (Carreira & Zisserman, 2017), while pose features are obtained using SignBERT (Hu et al., 2021b). We follow SEDS (Jiang et al., 2024a) for RGB–pose fusion and fine-grained retrieval matching, retaining at most 64 visual clips and 32 text tokens per sample. Frame-level gesture descriptions are generated offline using LLaVA-OneVision-7B (Li et al., 2025a) with a sign-language-oriented prompt and encoded using BERT-base-uncased. The SDE attention dimension is $d_a = 512$. We set $\gamma = 0.05$ for PHOENIX-2014T and 0.02 for How2Sign and CSL-Daily. The retrieval model is trained for 200 epochs using BertAdam with a batch size of 128, a base learning rate of 1×10^{-5}, a learning rate of 1×10^{-4} for the RGB projection branch, a weight decay of 0.2, and a warmup ratio of 0.1. After freezing the retrieval model, CRR is trained for 20 epochs using AdamW with a batch size of 128, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-4}. Its listwise encoder consists of one Transformer layer with four attention heads and a hidden dimension of 128. We set $K = 32$, $\alpha = 0.08$, $\tau = 0.07$, $\tau_r = 1.0$, and $\delta = 0.1$. The loss weights are $\lambda_{\text{ce}} = \lambda_{\text{mar}} = 1.0$ and $\lambda_{\text{rank}} = 0.25$.

Table 2: Ablation of the main components on CSL-Daily using RGB-only visual input.

SDE	CRR	T2V			V2T		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
		74.3	87.8	91.1	72.8	88.1	91.6
✓		76.4	89.7	92.7	75.9	88.6	92.1
	✓	78.1	88.3	91.1	73.5	88.6	91.8
✓	✓	**80.8**	**90.2**	**93.0**	**77.5**	**88.8**	**92.6**

Table 3: Comparison of visual–description fusion strategies on CSL-Daily.

Fusion Strategy	T2V R@1	V2T R@1	$mR@1$
None	74.3	72.8	73.5
Addition	75.7	73.8	74.7
Concatenation	73.7	71.7	72.7
Global Attention	71.2	70.8	71.0
Local Attention	**76.4**	**75.9**	**76.2**

4.2 Comparison with State-of-the-Art Methods

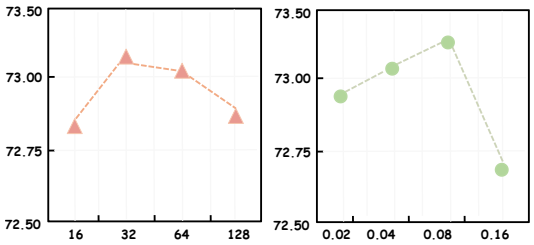
Evaluation on the PHOENIX-2014T dataset. As shown in Table 1, with RGB and pose inputs, SignRet improves T2V and V2T R@1 over SEDS* by 4.5 and 3.8 percentage points, reaching 79.9% and 80.1%, respectively. It also increases T2V R@5 from 91.6% to 92.2% and reduces V2T MnR from 3.1 to 3.0, while maintaining comparable R@10 performance.

Evaluation on the How2Sign dataset. SignRet improves T2V R@1, R@5, and R@10 over SEDS* by 2.0, 0.9, and 0.7 percentage points, respectively, while increasing V2T R@1 and R@5 by 1.9 and 0.7 points. T2V MnR is also reduced from 44.9 to 44.7, whereas V2T R@10 and MnR remain unchanged. This indicates improved top-ranked retrieval while maintaining performance.

Evaluation on the CSL-Daily dataset. Despite the strong SEDS* baseline, SignRet improves T2V R@1 from 84.5% to 88.7% and V2T R@1 from 84.1% to 87.0%, corresponding to gains of 4.2 and 2.9 percentage points. Improvements are also observed on R@5 and R@10, while MnR decreases from 3.3 to 3.0 for T2V and from 3.4 to 3.3 for V2T. These results confirm the effectiveness of SignRet even when the initial retrieval model is already strong.

4.3 Ablation Study

We evaluate the main components, fusion strategy, description granularity, visual inputs, and CRR hyperparameters. Unless otherwise specified, experiments use RGB-only inputs on CSL-Daily.



(a) Candidate pool size K (b) Residual weight α

Figure 3: Sensitivity analysis of K and α.

Effectiveness of the Main Components. Table 2 shows that SDE improves T2V/V2T R@1 from 74.3%/72.8% to 76.4%/75.9%, while CRR alone achieves 78.1%/73.5%. Combining both yields the best results of 80.8%/77.5%, with gains of 6.5/4.7 percentage points over the baseline, supporting their complementary contributions.

Feature Fusion Strategy. Table 3 compares different description fusion strategies on CSL-Daily under the RGB-only setting. Simple addition improves mR@1 from 73.5% to 74.7%, whereas concatenation and global attention perform worse than no fusion. Local attention achieves the best result of 76.2% mR@1. This result shows that temporally constrained local interaction is more effective than the compared direct and global fusion strategies.

Description Granularity. Table 4 evaluates different sign language description granularities on CSL-Daily. Video-level descriptions improve mR@1 from 73.5% to 75.1%, showing the benefit of auxiliary gesture descriptions. Frame-level descriptions achieve the best performance, reaching 76.2% mR@1. This result indicates that temporally localized descriptions capture more effective fine-grained gesture information than a single video-level description.

Table 4: Effect of sign language description granularity.

Granularity	T2V	V2T	$mR@1$
None	74.3	72.8	73.5
Video-level	75.6	74.6	75.1
Frame-level	**76.4**	**75.9**	**76.2**

Table 5: Effect of different visual inputs.

Input	T2V	V2T	$mR@1$
RGB	73.7	73.7	73.7
Pose	70.7	68.5	69.6
RGB + Pose	**79.9**	**80.1**	**80.0**

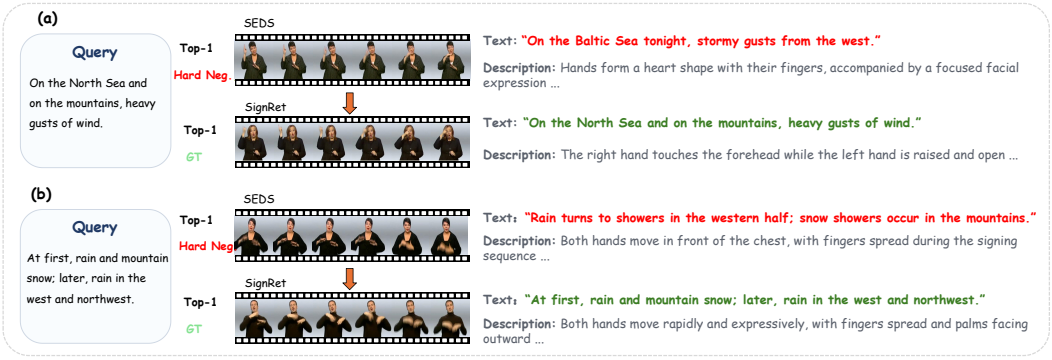


Figure 4: Qualitative text-to-video retrieval results on PHOENIX-2014T.

Effect of Visual Inputs. Table 5 compares different visual inputs on PHOENIX-2014T. RGB features outperform pose-only input, achieving 73.7% mR@1 compared with 69.6%. Combining RGB and pose yields the best performance, reaching 79.9% T2V R@1, 80.1% V2T R@1, and 80.0% mR@1. This corresponds to a 6.3 percentage-point improvement in mR@1 over RGB alone, demonstrating that appearance and pose information provide complementary cues.

Candidate Size K. Figure 3(a) shows that performance peaks at $K = 32$ and declines with larger candidate pools, suggesting that lower-ranked candidates may introduce noise into refinement. We therefore use $K = 32$.

Residual Weight α. Figure 3(b) shows that performance improves as α increases to 0.08, but drops at 0.16. This suggests that moderate corrections are beneficial, while excessive corrections may weaken the initial ranking. We set $\alpha = 0.08$.

4.4 Qualitative Results

Figure 4 presents two text-to-video retrieval examples from PHOENIX-2014T. In both cases, SEDS ranks a confusing hard negative first, whereas SignRet retrieves the ground-truth video at rank 1. The hard negatives share similar weather-related semantics and overall signing patterns with the queries, making them difficult to distinguish through pairwise similarity alone. The representative frame-level sign language descriptions expose differences in hand configurations, spatial positions, and inter-hand relations between the retrieved candidates, providing fine-grained gesture cues complementary to the RGB clip features.

5 Conclusion

In this work, we presented SignRet, a sign language retrieval framework that addresses insufficient fine-grained gesture representation and unreliable ranking among visually similar candidates. Sign Language Description Enhancement integrates MLLM-generated frame-level gesture descriptions into RGB clip features through temporally aligned local interaction, providing detailed gesture cues for cross-modal matching. Candidate Ranking Refinement treats the initially retrieved candidates as a structured list and performs listwise contextual reasoning, allowing each candidate to be evaluated relative to its competitors rather than independently. Experiments on three public datasets demonstrate consistent improvements in both text-to-video and video-to-text retrieval, validating the effectiveness of our framework.

AI USE STATEMENT

We used generative AI tools to assist with translation, drafting, and language editing of selected portions of the manuscript. We also sought feedback on the clarity and organization of the methodology section, including the terminology and explanations of model components. The authors reviewed and revised all AI-assisted text and checked the methodological descriptions against the actual implementation. We take full responsibility for the final content of this paper.

REFERENCES

Zechen Bai, Tianjun Xiao, Tong He, Pichao Wang, Zheng Zhang, Thomas Brox, and Mike Zheng Shou. Bridging information asymmetry in text-video retrieval: A data-centric approach. In *International Conference on Learning Representations*, volume 2025, pp. 43309–43331, 2025.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7784–7793, 2018b.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. C 2 rl: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19016–19026, 2023.

Cheol-Ho Cho, WonJun Moon, Woojin Jun, MinSeok Jung, and Jae-Pil Heo. Ambiguity-restrained text-video representation learning for partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2500–2508, 2025.

Luca Collorone, Matteo Gioia, Massimiliano Pappa, Paolo Leoni, Giovanni Ficarra, Or Litany, Indro Spinelli, and Fabio Galasso. Monster: a unified model for motion, scene, text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10940–10949, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2735–2744, 2021.

Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14094–14104, 2022.

Mo Guan, Yan Wang, Guangkun Ma, Jiarui Liu, and Mingzu Sun. Mska: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recognition*, 165:111602, 2025.

Bowen Guo, Shiwei Gan, Yafeng Yin, Xiao Liu, Zhiwei Jiang, and Shunmei Meng. Sentence-level segmentation for long sign language videos with captions. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3428–3437, 2025.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11087–11096, October 2021a.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11087–11096, 2021b.

Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park, and Hyeyoung Park. Narrating the video: boosting text-video retrieval via comprehensive utilization of frame-level captions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24077–24086, 2025.

Junkyu Jang. Corecaption: Core caption based text-to-video retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6600–6610, 2026.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8742–8752, 2025.

Longtao Jiang, Min Wang, Zecheng Li, Yao Fang, Wengang Zhou, and Houqiang Li. Seds: Semantically enhanced dual-stream encoder for sign language retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5141–5150, 2024a.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. Signclip: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9171–9193, 2024b.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. Leveraging the power of mllms for gloss-free sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21048–21058, October 2025.

Junmyeong Lee, Chan Hur, ChangSu Choi, Sukmin Cho, Fitsum Gaim, Eui Jun Hwang, Hoyun Song, and KyungTae Lim. Semantic hardness is not visual hardness: Sign-aware hard negative mining for sign language retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28262–28277, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*, 2025b.

Xiao Liu, Shiwei Gan, Yafeng Yin, Bowen Guo, Zhiwei Jiang, Shunmei Meng, Lei Xie, and Sanglu Lu. Signpr: A progressive vector-quantized diffusion framework for sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2198–2208, 2026.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M De Melo, Benjamin Van Durme, and Rama Chellappa. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19691–19701, 2025.

Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Yongjun Bao, Guiguang Ding, et al. Tempme: Video temporal token merging for efficient text-video retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 60839–60860, 2025.

Lucas Ventura, Cordelia Schmid, and Gül Varol. Learning text-to-video retrieval from image captioning. *International Journal of Computer Vision*, 133(4):1834–1854, 2025.

Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 12–20, 2019.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2816–2827, 2023.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10704–10713, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pp. 53285–53310, 2025.

Zhiqian Zhao, Liang Li, Lei Shen, Xichun Sheng, Yaoqi Sun, Fang Kang, and Chenggang Yan. Temporal calibrating and distilling for scene-text aware text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13323–13331, 2026.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1316–1325, 2021.

A Appendix

This appendix provides additional details and results for SignRet. Section B describes the generation of frame-level gesture descriptions. Section C presents component ablations with RGB-only and RGB–pose inputs. Section D provides successful and failed retrieval examples in both directions. Section E reports retrieval-stage latency and model parameter counts.

B Construction of Sign Language Descriptions

Frame-level sign language descriptions provide a complementary view of locally observable gesture states. Unlike the paired retrieval sentence, which summarizes the meaning conveyed by the complete video, these descriptions characterize fine-grained hand configurations, spatial locations, body postures, facial expressions, and inter-hand relations at individual time steps. Their temporal sequence further reflects how these gesture states evolve throughout the video. For each video, the descriptions are generated offline and encoded using BERT-base-uncased. The resulting description features are temporally aligned with the corresponding RGB clips before visual–description interaction.

B.1 Frame-Level Description Generation

Following the frame-level description generation setting of MMSLT (Kim et al., 2025), we use LLaVA-OneVision-7B (Li et al., 2025a) with the same prompt and generation settings to generate a natural-language description for each sampled video frame. Given a sign language video V_i containing T sampled frames,

$$V_i = \{I_{i,t}\}_{t=1}^{T}, \tag{12}$$

we independently feed each frame $I_{i,t}$ into LLaVA-OneVision-7B and generate its frame-level description as

$$D_{i,t} = f_{\text{LLaVA-OneVision}}\left(I_{i,t}, \mathcal{P}_{\text{desc}}\right), \tag{13}$$

where $\mathcal{P}_{\text{desc}}$ is the prompt "Describe only the motion and gestures of the person in the image, focus on hands and face." The prompt focuses the model on visually observable signing cues, including handshape, hand location, hand orientation, body posture, gaze, mouth shape, and facial expression, while suppressing irrelevant information about the signer's appearance, clothing, and background. Applying this procedure to all sampled frames produces the description sequence

$$\mathcal{D}_i = [D_{i,1}, D_{i,2}, \ldots, D_{i,T}]\,. \tag{14}$$

The paired retrieval sentences and gloss annotations are not provided to the MLLM during generation. Thus, the descriptions supply complementary fine-grained gesture information without directly revealing the retrieval target. All descriptions are generated before retrieval-model training; therefore, LLaVA-OneVision-7B is not required during training or inference. The descriptions are subsequently encoded using BERT-base-uncased (Devlin et al., 2019) and temporally aligned with the corresponding RGB clips.

C Additional Ablation Studies

SDE	CRR	T2V			V2T		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
		68.4	87.1	91.3	69.3	86.9	91.6
✓		69.9	87.1	**91.9**	71.2	87.9	**92.4**
	✓	72.6	**88.9**	91.7	70.7	**88.8**	91.7
✓	✓	**73.7**	88.0	**91.9**	**73.7**	88.3	**92.4**

Table 6: Ablation of the main components on PHOENIX-2014T. All variants use RGB features only. SDE and CRR denote Sign Language Description Enhancement and Candidate Ranking Refinement, respectively.

SDE	CRR	T2V			V2T		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
		74.8	90.3	95.5	75.7	90.7	94.7
✓		75.7	91.1	95.5	77.1	91.6	95.2
✓	✓	**79.9**	**92.2**	**95.6**	**80.1**	**92.5**	**95.5**

Table 7: Ablation of the main components on PHOENIX-2014T. All variants use both RGB and pose features. The baseline and its variants are trained under the same setting. SDE and CRR denote Sign Language Description Enhancement and Candidate Ranking Refinement, respectively.

Component ablation with RGB features. Table 6 evaluates the individual and combined effects of SDE and CRR on PHOENIX-2014T (Camgoz et al., 2018a) using only RGB features. Adding SDE improves the T2V and V2T R@1 scores from 68.4 and 69.3 to 69.9 and 71.2, respectively, while CRR increases them to 72.6 and 70.7. Combining SDE with CRR achieves the best R@1 scores of 73.7 in both retrieval directions, corresponding to improvements of 5.3 and 4.4 percentage points over the RGB-only baseline. These results show that SDE and CRR provide complementary benefits for sign language retrieval.

Component ablation with RGB and pose features. Table 7 further evaluates the proposed components on PHOENIX-2014T using both RGB and pose features (Hu et al., 2021a). Following the retrieval backbone, the pose features are processed by a separate pose branch and fused with the RGB representations. SDE enhances only the RGB branch through local visual–description attention, while the pose branch remains unchanged. Unlike SEDS* in the main comparison, which is evaluated using the released checkpoint, the RGB+Pose baseline in this ablation is retrained under the same experimental setting as its variants. Adding SDE improves T2V and V2T R@1 from 74.8 and 75.7 to 75.7 and 77.1, respectively. Combining SDE with CRR further increases the scores to 79.9 and 80.1, outperforming the retrained RGB+Pose baseline by 5.1 and 4.4 percentage points. These results show that the description cues remain complementary to the structural information captured by pose features.

D Additional Qualitative Results

We provide additional qualitative results for both text-to-video (T2V) and video-to-text (V2T) retrieval. These examples illustrate how frame-level sign language descriptions and CRR contribute to the final retrieval results, as well as the remaining limitations of the proposed framework.

Successful cases. Figures 5(a) and 6(a) present two examples in which the full model successfully corrects the initial retrieval errors. In both cases, the initial results are difficult to distinguish using pairwise matching scores alone. In the T2V example, the incorrect and ground-truth videos exhibit similar overall signing patterns, whereas in the V2T example, the retrieved and ground-truth sentences express closely related weather semantics. Frame-level sign language descriptions provide fine-grained gesture evidence, including hand configurations, relative hand positions, and facial expressions. By combining description-enhanced representations with listwise candidate refinement, SignRet promotes the ground-truth result to the first position.

Failure cases. Figures 5(b) and 6(b) present two particularly challenging cases for the full model. In the T2V example, the incorrectly retrieved and ground-truth videos contain similar signing patterns, whereas in the V2T example, the two candidate sentences contain overlapping weather conditions and geographic references. The decisive distinctions are not sufficiently captured by the description-enhanced representations or the initial score distribution, preventing CRR from correcting the top-ranked result. These cases reveal a limitation of the current framework when retrieval depends on subtle temporal dynamics or highly similar sentence semantics.

SDE	CRR	Retrieval Time (ms/query)		Params (M)
		T2V	V2T	
		1.89	1.87	303.8
✓		2.51	2.50	312.8
✓	✓	4.50	5.15	333.6

Table 8: Efficiency analysis of the main components on PHOENIX-2014T. Retrieval time denotes the average latency per query from cached representations to the final retrieval scores, while Params denotes the number of online model parameters. The offline LLaVA-OneVision-7B and the description encoder are excluded.

E Complexity Analysis

The sign language descriptions used by SDE are generated offline. Therefore, LLaVA-OneVision-7B and the description encoder are excluded from the online retrieval pipeline. Table 8 reports the average retrieval-stage latency per query and the online model parameters on PHOENIX-2014T. Runtime is measured on a single NVIDIA RTX 4090 from cached text, pose, RGB, and description representations to the final retrieval scores. The gallery contains 642 videos, and the query batch size and gallery chunk size are both set to 64. All runtime results are averaged over three runs. Adding SDE increases the T2V/V2T latency from 1.89/1.87 ms to 2.51/2.50 ms, introducing only a modest overhead. After further applying CRR, the complete model requires 4.50 ms for T2V and 5.15 ms for V2T, as CRR performs listwise refinement over the ranking descriptors of the selected candidates. The parameter count increases from 303.8M for the baseline to 312.8M with SDE and 333.6M for the complete SDE+CRR model. These counts cover the online retrieval and refinement networks without counting shared backbone parameters more than once.

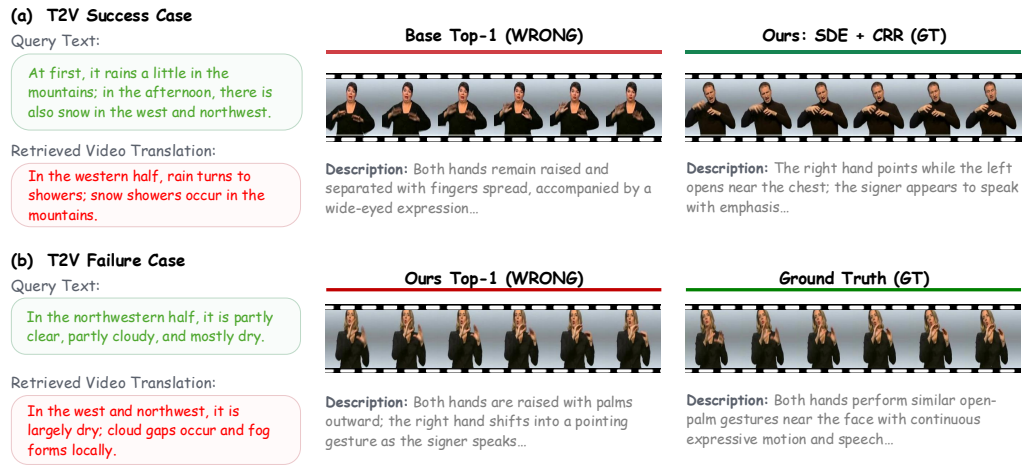


Figure 5: Qualitative examples of text-to-video (T2V) retrieval. (a) A success case where our method retrieves the ground-truth sign video while the baseline returns an incorrect result. (b) A failure case where the top-1 video retrieved by our method differs from the ground truth.

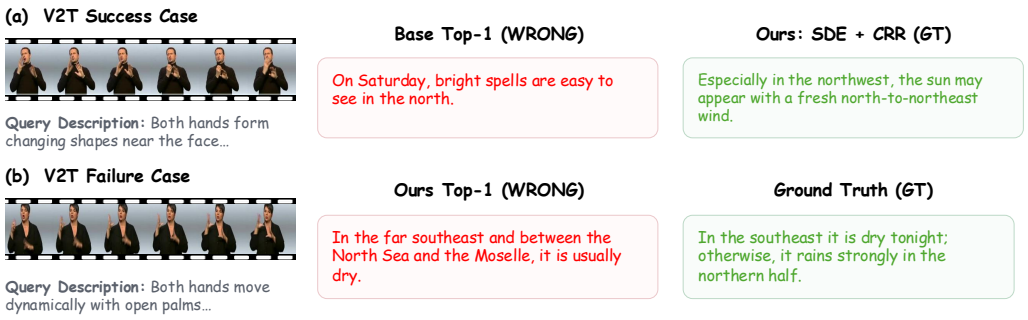


Figure 6: Qualitative examples of video-to-text (V2T) retrieval. (a) A success case where our method retrieves the ground-truth text while the baseline produces an incorrect result. (b) A failure case where the top-1 text retrieved by our method differs from the ground truth.