SIGNALIGN: STRUCTURE-AWARE ALIGNMENT FOR CROSS-MODAL REPRESENTATION LEARNING IN SIGN LANGUAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign language retrieval enables bidirectional search between natural-language text and signing videos based on their semantic correspondence. Existing methods compute video–text similarity by comparing global representations or aggregating local similarities between video tokens and text tokens. Despite their effectiveness, two limitations remain. First, contrastive training emphasizes annotated video–text associations without explicitly preserving semantic relations among texts. As text representations adapt to videos, their pretrained relative distances can change. If texts with different meanings become too close, a video associated with one may be incorrectly retrieved for the other. retrieved for the other. Second, local alignment computes weights independently for each text token. A video token with high similarity to several text tokens can increase the retrieval score through all of them, even when the video does not express the remaining content of the text. To address these limitations, we propose SignAlign, a structure-aware alignment framework for cross-modal representation learning. It preserves semantic relations among sentences while jointly aligning video tokens and text tokens. To preserve sentence relations during adaptation, Text Semantic Structure Alignment (TSSA) uses a frozen pretrained text encoder as a reference. It aligns normalized pairwise distances and triplet-wise angles among sentence embeddings with those derived from this reference. Complementing this sentence-level alignment, Cross-Modal Token Structure Alignment (CTSA) jointly determines token alignment weights through entropy-regularized optimal transport. Text tokens aligned with the same video token share its available weight, limiting repeated emphasis in retrieval scoring. The resulting joint alignment score is fused with independently weighted local similarity scores during training and inference. Experiments on three public sign language retrieval benchmarks demonstrate improved retrieval in both directions over strong baselines.

1 INTRODUCTION

Sign language retrieval (Albanie et al., 2020; Varol et al., 2021; Zuo et al., 2025; Yao et al., 2026; Zhao et al., 2026a;b) enables bidirectional search between natural-language text and signing videos: a text query retrieves videos conveying the corresponding meaning, while a video query retrieves corresponding texts. By connecting videos with corresponding texts, it facilitates access to and organization of sign language resources in education, public services, and online media. Users can locate relevant videos through text or retrieve corresponding textual information from a video, reducing the effort required to browse and organize these materials manually. However, matching text with signing videos is challenging because meaning is conveyed through the coordinated use of hand movements, facial expressions, and body posture over time (Yin et al., 2021; Zhou et al., 2020; Chen et al., 2022; Zhang et al., 2026b; Bian et al., 2026). The relevant information can span multiple video segments and involve several visual cues acting together. Retrieval models must therefore integrate these cues across the video and relate their combined meaning to the text, while retaining the details that distinguish different utterances.

Existing methods (Radford et al., 2021; Jia et al., 2021; Ye et al., 2024; Wong et al., 2025; Li et al., 2025; Kim et al., 2025; Asasi et al., 2026; Madjoukeng et al., 2025; Yazdani et al., 2025; Keleş et al.,

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.

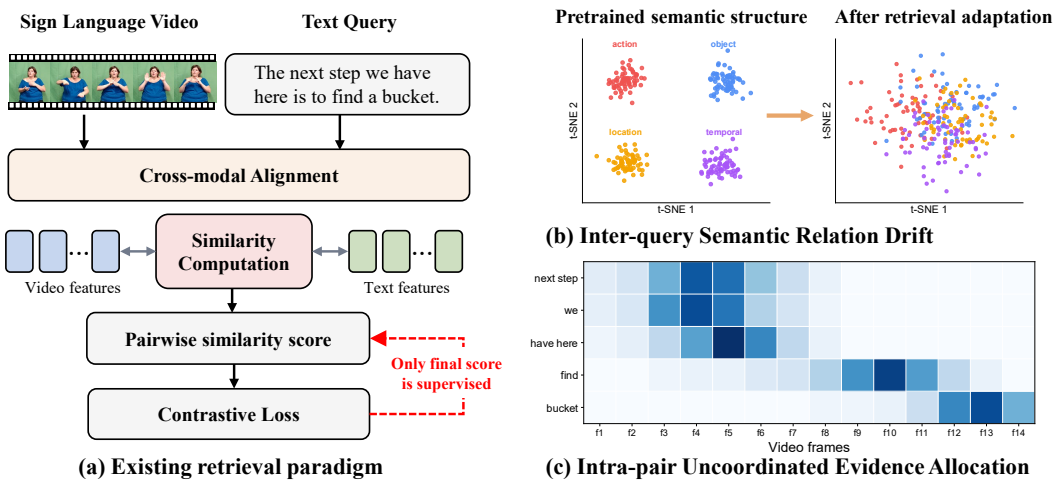


Figure 1: Motivation for structure-aware alignment in sign language retrieval. (a) Contrastive retrieval training supervises video–text similarity scores. (b) Retrieval adaptation can change the relative geometry of sentence embeddings. (c) Multiple text tokens repeatedly rely on the same few video segments for matching.

2026; Jiang et al., 2024b) learn video and text representations in a shared embedding space and rank candidates by their similarity. Duarte et al. (Duarte et al., 2022) compare global representations of videos and texts, whereas CiCo and SEDS aggregate similarities between individual text tokens and video segments into retrieval scores (Cheng et al., 2023; Jiang et al., 2024a). UPRet further introduces probabilistic representations and transport-based local alignment during training (Wu et al., 2024). These methods use contrastive supervision (Cheng et al., 2023; Jiang et al., 2024a; Tan et al., 2025; Zhang et al., 2026a) to increase the similarity of annotated video–text matches relative to other candidate combinations, as illustrated in Fig. 1(a). However, this supervision does not explicitly preserve semantic relationships among texts during adaptation, potentially making texts with different meanings harder to distinguish when matching them to videos. For local scoring methods that normalize segment weights independently for each text token, several tokens can also assign high weights to the same segment. A few segments can therefore dominate the retrieval score, allowing a video to rank highly even when it matches only part of the text.

These limitations can undermine retrieval by weakening semantic distinctions between texts and allowing partial visual matches to dominate the score. At the sentence level, if texts with different meanings have similar representations, a video corresponding to one may receive a high score for the other. Pretrained sentence embeddings encode semantic relationships that provide a reference for these distinctions (Gao et al., 2021; Hamidullah et al., 2026; Wong et al., 2024). During retrieval adaptation, however, contrastive training updates the text encoder to favor annotated video–text associations over other candidates, including semantically related texts. Without explicit constraints on text–text relationships, these updates can alter the pretrained sentence geometry (Moon et al., 2025; Guo et al., 2025), potentially weakening distinctions needed to identify semantically correct candidates, as illustrated in Fig. 1(b). At the token level, alignment weights determine which similarities between video and text tokens most strongly affect the retrieval score. When these weights are normalized separately for each text token, a video token similar to several text tokens can be emphasized repeatedly, as illustrated in Fig. 1(c). This repeated emphasis can inflate the score based on a few locally similar features while obscuring mismatches with the rest of the sentence. These issues motivate preserving pretrained sentence relationships during adaptation and jointly determining alignment weights across video and text tokens.

To this end, we propose SignAlign, a structure-aware alignment framework for cross-modal representation learning in sign language retrieval. At the sentence level, Text Semantic Structure Alignment (TSSA) preserves pretrained relationships among sentence embeddings during retrieval adaptation. It aligns their normalized pairwise distances and triplet-wise angles with those derived from a frozen pretrained text encoder, helping retain semantic distinctions among texts. At the token level, Cross-Modal Token Structure Alignment (CTSA) jointly determines alignment weights between video and text tokens through entropy-regularized optimal transport. Its constraints limit the

weight that a video token can accumulate across text tokens, reducing repeated emphasis on the same visual content. CTSA uses these jointly determined weights to aggregate local similarities into a video–text alignment score. This score is fused with the local aggregation scores computed using independently normalized alignment weights. The fusion retains strong local correspondences while incorporating joint alignment across tokens, reducing the risk that a few local similarities dominate the scores used for bidirectional retrieval. Our contributions are summarized as follows:

- We propose SignAlign, a structure-aware alignment framework for sign language retrieval that jointly models relations among sentence representations and assigning weights within video–text pairs.
- We introduce TSSA to preserve pretrained semantic geometry during retrieval adaptation. It aligns normalized pairwise distances and triplet-wise angles among sentence embeddings with those derived from a frozen pretrained text encoder.
- We develop CTSA to jointly assign token-to-segment matching weights. It constrains the total matching weight assigned to each text token and video segment. The resulting transport score is blended with the original retrieval scores, which aggregate local token-to-segment similarities.
- We evaluate SignAlign on PHOENIX-2014T, CSL-Daily, and How2Sign, demonstrating improved bidirectional retrieval performance over baselines.

2 Related Work

Sign Language Retrieval. Sign language retrieval learns video–text correspondence for bidirectional search. Early work compares global video and text embeddings and improves video features through iterative sign spotting and alignment (Duarte et al., 2022). CiCo (Cheng et al., 2023) combines domain-aware sign encoding with cross-lingual contrastive learning. SEDS (Jiang et al., 2024a) integrates RGB and pose cues through semantic aggregation and dual-stream matching, while SignCLIP (Jiang et al., 2024b) learns transferable representations from multilingual sign-text data. UPRet (Wu et al., 2024) models correspondence uncertainty through probabilistic representations. SAN (Lee et al., 2026) constructs hard negative captions through word substitution guided by visually confusable signs. These methods emphasize video–text associations, with limited explicit preservation of semantic relations among texts. SignAlign combines such preservation through TSSA with joint alignment between video tokens and text tokens through CTSA.

Structure-Aware Cross-Modal Alignment. Relational knowledge distillation (Park et al., 2019) transfers pairwise distances and triplet-wise angles, while similarity-preserving distillation (Tung & Mori, 2019) preserves pairwise similarity patterns. TCPL (Moon et al., 2025) applies distance and angular supervision to preserve CLIP text relations in partially relevant video retrieval. For local alignment, FILIP (Yao et al., 2022) uses token-wise maximum similarity, X-CLIP (Ma et al., 2022) aggregates multi-granular similarities through attention, and S-ViLM (Xiong et al., 2024) incorporates temporal grouping and spatial grounding. Transport-based methods jointly determine local associations. UNITER (Chen et al., 2020) uses word-region transport for pretraining, while Norton (Lin et al., 2024) addresses noisy clip-caption correspondence in long videos. UPRet (Wu et al., 2024) introduces frame-word transport within a probabilistic framework as a training objective. SignAlign couples sentence geometry preservation through TSSA with joint token alignment through CTSA. CTSA incorporates jointly constrained alignments into retrieval scoring, limiting repeated emphasis on a few video tokens while retaining strong local similarities.

3 Method

3.1 Problem Formulation

Sign language retrieval aims to retrieve semantically corresponding videos from textual queries and texts from video queries. Given a mini-batch of N paired examples $\{(v_i, t_i)\}_{i=1}^{N}$, where v_i denotes a sign language video and t_i its paired text, the objective is to learn a similarity function $s(v_i, t_j)$

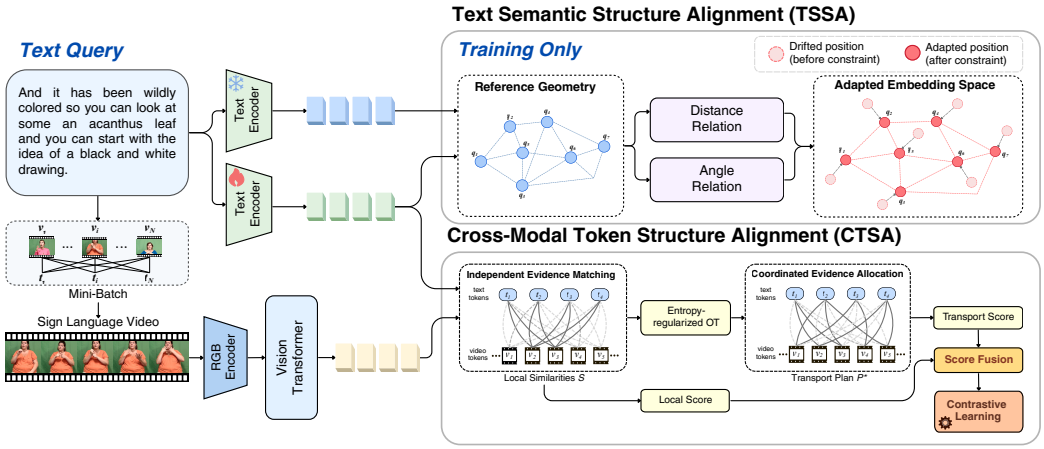


Figure 2: Overall framework of our proposed SignAlign.

that assigns higher scores to matched pairs than to mismatched ones. For text-to-video (T2V) retrieval, candidate videos are ranked by $s(v_i, t_j)$ for a fixed textual query t_j. For video-to-text (V2T) retrieval, candidate texts are ranked using the same function for a fixed video query v_i.

3.2 Overview of SignAlign

Sign language retrieval ranks candidate videos or texts according to their cross-modal similarity. Figure 2 illustrates how SignAlign computes these similarities. Given a video and a text, an RGB encoder followed by a vision Transformer produces visual tokens, while a text encoder produces textual representations. Both modalities are projected into a shared embedding space to compare text tokens with video segments. During training, Text Semantic Structure Alignment (TSSA) aligns relationships among sentence representations with those from a frozen pretrained text encoder. This preserves semantic distinctions among texts as their representations adapt to signing videos. Using the token-to-segment similarities, Cross-Modal Token Structure Alignment (CTSA) determines matching weights jointly through optimal transport, limiting repeated emphasis on the same video segments. These weights are used to compute a video–text similarity score. The score fusion block combines this score with the original retrieval scores obtained by independently aggregating local token-to-segment similarities. The blended scores supervise retrieval training and rank candidates at inference, returning the most relevant videos or texts.

3.3 Text Semantic Structure Alignment

Contrastive training favors a video's annotated text over other texts without explicitly preserving their semantic relationships (Moon et al., 2025). If semantically different texts become too close during adaptation, a video may be retrieved for the wrong text. TSSA preserves pretrained relative distances and angles among sentence embeddings to retain these semantic distinctions.

Reference Semantic Geometry. As illustrated in the upper branch of Fig. 2, a frozen copy of the pretrained text encoder provides the reference geometry. Given the texts $\{t_i\}_{i=1}^{N}$ in a mini-batch, let $\mathbf{z}_i = f_t(t_i)$ and $\bar{\mathbf{z}}_i = \bar{f}_t(t_i)$ denote their projected EOT embeddings from the trainable and frozen encoders, respectively. Both encoders receive the same texts. TSSA constrains the relations among the trainable embeddings using those among the frozen embeddings, without requiring the two representations of each sentence to coincide. This allows the sentence representations to adapt to retrieval while retaining their relative semantic organization.

Distance Relation Alignment. Following relational knowledge distillation (Park et al., 2019), we constrain the relative distances between sentence embeddings. For sentences t_i and t_j, let $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$ and $\bar{d}_{ij} = \|\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j\|_2$ denote their Euclidean distances in the trainable and frozen embedding spaces, respectively. Here, $\|\cdot\|_2$ denotes the Euclidean norm. We normalize each distance by the mean of all nonzero pairwise distances within the same mini-batch and embedding space. Denoting these means by μ_d and $\bar{\mu}_d$, respectively, the normalized distances are $\tilde{d}_{ij} = d_{ij}/\mu_d$

and $\tilde{\bar{d}}_{ij} = \bar{d}_{ij}/\bar{\mu}_d$. This normalization compares relative sentence separation without requiring the two spaces to have the same overall scale. The distance loss penalizes the difference between the normalized distances:

$$\mathcal{L}_D = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \rho\Big(\tilde{d}_{ij} - \tilde{\bar{d}}_{ij}\Big), \qquad \rho(e) = \begin{cases} \frac{1}{2}e^2, & |e| < 1, \\ |e| - \frac{1}{2}, & |e| \geq 1. \end{cases} \tag{1}$$

Here, $\mathcal{P} = \{(i,j) : 1 \leq i,j \leq N,\ i \neq j\}$ is the set of ordered pairs of distinct batch indices, and $|\mathcal{P}| = N(N-1)$ is its size. The function ρ is the Smooth L1 penalty, with $e = \tilde{d}_{ij} - \tilde{\bar{d}}_{ij}$ denoting the normalized distance discrepancy. It penalizes small discrepancies quadratically and large discrepancies linearly, limiting the influence of large deviations. The resulting loss aligns normalized pairwise distances among trainable sentence embeddings with those in the frozen reference space.

Angular Relation Alignment. In addition to pairwise distances, TSSA explicitly supervises the angle formed by three sentence embeddings. For a triplet (i,j,k), the cosine of the angle at $\mathbf{z}_i$ is

$$a_{ijk} = \frac{(\mathbf{z}_j - \mathbf{z}_i)^\top (\mathbf{z}_k - \mathbf{z}_i)}{\|\mathbf{z}_j - \mathbf{z}_i\|_2 \|\mathbf{z}_k - \mathbf{z}_i\|_2}. \tag{2}$$

The corresponding reference cosine $\bar{a}_{ijk}$ is defined using the frozen embeddings. The angular loss penalizes discrepancies between these quantities:

$$\mathcal{L}_A = \frac{1}{|\mathcal{T}|} \sum_{(i,j,k)\in\mathcal{T}} \rho(a_{ijk} - \bar{a}_{ijk}), \tag{3}$$

where $\mathcal{T}$ contains triplets with distinct indices and nonzero difference vectors in both spaces. This loss aligns triplet-wise angular relations with those in the frozen reference space, complementing pairwise distance alignment.

The two alignment losses are combined as $\mathcal{L}_{\text{TSSA}} = \lambda_d \mathcal{L}_D + \lambda_a \mathcal{L}_A$, where λ_d and λ_a weight distance and angular alignment, respectively.

3.4 Cross-Modal Token Structure Alignment

Independent local matching may rely on the same few video segments to match different text tokens. These repeated matches can produce a high retrieval score even when other parts of the text are poorly matched. CTSA jointly determines token–segment associations to limit repeated emphasis on the same segments when computing video–text similarity.

Cross-modal Token Similarity. For each candidate video–text pair (v_i, t_j), let $\mathbf{v}_f$ denote the embedding of video segment f and $\mathbf{t}_\ell$ the embedding of text ℓ in the shared embedding space. Here, $f \in \{1,\ldots,F\}$ and $\ell \in \{1,\ldots,L\}$ index the video and text tokens used for matching, respectively, with sample indices omitted for clarity. As illustrated in the lower branch of Fig. 2, their pairwise cosine similarities form a matrix $M \in \mathbb{R}^{F\times L}$, where $M_{f\ell} = \cos(\mathbf{v}_f, \mathbf{t}_\ell)$. Each row corresponds to a video segment and each column to a word.

Token-level Structure Alignment. Given the similarity matrix M, CTSA jointly determines the similarity structure between text tokens and video tokens. These weights form a transport matrix $P \in \mathbb{R}_+^{F\times L}$, where $P_{f\ell}$ is the matching weight between segment f and word ℓ. We formulate this assignment as entropy-regularized optimal transport (Cuturi, 2013):

$$P^\star = \arg\max_{P\geq 0} \left[\sum_{f=1}^{F}\sum_{\ell=1}^{L} P_{f\ell} M_{f\ell} + \tau H(P) \right], \quad \text{s.t.} \quad \sum_{\ell=1}^{L} P_{f\ell} = \frac{1}{F}\ \forall f, \quad \sum_{f=1}^{F} P_{f\ell} = \frac{1}{L}\ \forall \ell. \tag{4}$$

Here, $P^\star$ contains the optimal matching weights between video tokens and text tokens. The similarity term favors token pairs with higher cosine similarity. The entropy term $H(P) = -\sum_{f=1}^{F}\sum_{\ell=1}^{L} P_{f\ell}\log P_{f\ell}$ discourages concentrating weights on only a few token pairs, with larger τ producing more evenly distributed weights. Under the shared constraints, different text tokens must share a video token's matching weight, limiting repeated emphasis on the same video tokens.

These constraints require different text tokens to share the matching weight assigned to a video token. This limits repeated emphasis on a few video tokens, which could otherwise produce a high retrieval score despite matching only part of the text. We obtain an approximate transport matrix $\widehat{P}$ using a fixed number of log-domain Sinkhorn iterations.

Alignment-Based Retrieval Scoring. The joint alignment weights in $\widehat{P}$ define a similarity score for the candidate video v_i and text t_j:

$$s_{\mathrm{CTSA}}(v_i, t_j) = \sum_{f=1}^{F} \sum_{\ell=1}^{L} \widehat{P}_{f\ell} M_{f\ell}. \tag{5}$$

The entropy term affects the alignment weights but is not included in this score.

In parallel, the two local aggregation branches of CiCo (Cheng et al., 2023) compute the original retrieval scores from local similarities between video tokens and text tokens. We denote these scores by $s^{(r)}_{\mathrm{native}}(v_i, t_j)$, where $r \in \{1, 2\}$ indexes the aggregation branch. Each original score is blended with the same joint alignment score:

$$s^{(r)}(v_i, t_j) = (1 - \beta) s^{(r)}_{\mathrm{native}}(v_i, t_j) + \beta s_{\mathrm{CTSA}}(v_i, t_j), \tag{6}$$

where $\beta \in [0, 1]$ balances the original retrieval score and the joint alignment score.

3.5 Optimization with Combined Alignment

We retain the two local similarity aggregation rules of CiCo (Cheng et al., 2023). For $r = 1$, similarities are aggregated over text tokens and then averaged over video tokens. For $r = 2$, similarities are aggregated over video tokens and then averaged over text tokens. Each resulting local alignment score is fused with the same CTSA alignment score. Let $s^{(r)}_{ij} = s^{(r)}(v_i, t_j)$ denote the fused score for video v_i and text t_j under aggregation rule $r \in \{1, 2\}$. Each fused score matrix receives contrastive supervision in both retrieval directions:

$$\mathcal{L}^{(r)}_{\mathrm{V2T}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\gamma s^{(r)}_{ii})}{\sum_{j=1}^{N} \exp(\gamma s^{(r)}_{ij})}, \qquad \mathcal{L}^{(r)}_{\mathrm{T2V}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\gamma s^{(r)}_{ii})}{\sum_{j=1}^{N} \exp(\gamma s^{(r)}_{ji})}, \tag{7}$$

where N is the mini-batch size and $\gamma > 0$ is the similarity scaling factor. The retrieval objective averages the losses over both aggregation rules and retrieval directions:

$$\mathcal{L}_{\mathrm{RET}} = \frac{1}{4} \sum_{r=1}^{2} \left(\mathcal{L}^{(r)}_{\mathrm{V2T}} + \mathcal{L}^{(r)}_{\mathrm{T2V}} \right). \tag{8}$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{RET}} + \mathcal{L}_{\mathrm{TSSA}}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SignAlign on three widely used sign language benchmarks spanning different sign languages and content domains. **PHOENIX-2014T** (Camgoz et al., 2018) contains German Sign Language videos from weather forecast broadcasts, paired with German sentence-level translations. **CSL-Daily** (Zhou et al., 2021) consists of Chinese Sign Language videos and corresponding Chinese translations, covering daily-life topics such as travel, shopping, and medical care. **How2Sign** (Duarte et al., 2021) provides studio-recorded American Sign Language videos based on English instructional content, together with corresponding English transcripts.

Evaluation Metrics. Following previous sign language retrieval studies (Cheng et al., 2023), we evaluate both text-to-video (T2V) and video-to-text (V2T) retrieval using Recall@K ($K = 1, 5, 10$), Median Rank (MedR), and Mean Rank (MnR). Higher Recall@K and lower MedR and MnR indicate better performance. We additionally report $mR@1$, the average Recall@1 over the two retrieval directions, as the primary metric for overall comparison.

Table 1: Comparison with state-of-the-art methods on PHOENIX-2014T. † denotes our RGB-only reproduction of SEDS. Bold indicates the best results.

Method	Text-to-Video					Video-to-Text					mR@1↑
	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	
SA-SR (Duarte et al., 2022)	30.2	53.1	63.4	4.5	–	28.8	52.0	60.8	–	–	29.5
SA-CM (Duarte et al., 2022)	48.6	76.5	84.6	2.0	–	50.3	78.4	84.4	**1.0**	–	49.5
SA-COMB (Duarte et al., 2022)	55.8	79.6	87.2	**1.0**	–	53.1	79.4	86.1	**1.0**	–	54.5
GFSLT-VLP (Lee et al., 2026)	67.9	88.4	93.8	–	–	69.4	88.7	93.3	–	–	68.7
CiCo (Cheng et al., 2023)	69.5	86.6	92.1	**1.0**	–	70.2	88.0	92.8	**1.0**	–	69.9
SEDS-RGB† (Jiang et al., 2024a)	70.4	88.6	94.4	**1.0**	4.4	70.1	88.8	**94.5**	**1.0**	4.6	70.3
UPRet (Wu et al., 2024)	72.0	89.1	94.1	**1.0**	4.4	72.0	89.4	93.3	**1.0**	4.6	72.0
SignAlign (Ours)	**74.0**	**89.6**	**94.5**	**1.0**	**4.2**	**76.0**	**91.1**	93.2	**1.0**	**4.4**	**75.0**

Table 2: Comparison with existing methods on CSL-Daily under the RGB-only setting. † denotes our RGB-only reproduction of SEDS.

Method	Text-to-Video					Video-to-Text					$mR@1$
	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	
CiCo (Cheng et al., 2023)	75.3	88.2	91.9	1.0	-	74.7	89.4	92.2	1.0	-	75.0
SEDS-RGB† (Jiang et al., 2024a)	75.4	88.3	92.4	1.0	9.0	73.5	87.7	92.3	1.0	7.2	74.5
UPRet (Wu et al., 2024)	78.4	89.1	92.0	1.0	6.7	77.0	89.2	92.7	1.0	5.5	77.7
SignAlign (Ours)	**84.1**	**92.4**	**94.7**	**1.0**	**4.3**	**80.4**	**92.8**	**95.2**	**1.0**	**3.6**	**82.3**

Implementation Details. All experiments are conducted on a NVIDIA A800 GPU under the RGB-only setting. We describe the configuration for PHOENIX-2014T below. We use pre-extracted 1024-dimensional RGB features and uniformly sample or pad each video to 64 visual tokens. The text encoder is initialized from the text tower of CLIP ViT-B/32, and text and video representations are projected into a shared 512-dimensional embedding space. Following CiCo, we train for 200 epochs with a batch size of 128 and random seed 42. We use BertAdam with a weight decay of 10^{-3}, a warm-up ratio of 0.1, and a cosine learning rate schedule. The learning rate is set to 10^{-5} for both encoders. For TSSA, we set the distance and angular loss weights to $\lambda_d = 7.5$ and $\lambda_a = 30$, respectively. For CTSA, we use a blending coefficient of $\beta = 0.075$, a transport temperature of $\tau = 0.01$, and 10 Sinkhorn iterations. *Detailed configurations are provided in the appendix.*

4.2 Comparison with State-of-the-Art Methods

Evaluation on PHOENIX-2014T. As shown in Table 1, SignAlign achieves the highest $mR@1$ of 75.0%, with T2V and V2T R@1 scores of 74.0% and 76.0%, respectively. Compared with UPRet, the strongest baseline by $mR@1$, these results represent gains of 2.0, 4.0, and 3.0 percentage points in T2V R@1, V2T R@1, and $mR@1$, respectively.

Evaluation on CSL-Daily. SignAlign reaches an $mR@1$ of 82.3% on CSL-Daily (Table 2), exceeding UPRet by 4.6 percentage points. Its T2V and V2T R@1 scores of 84.1% and 80.4% improve over UPRet by 5.7 and 3.4 percentage points, respectively. It also achieves the highest R@5 and R@10 and the lowest reported MnR in both directions.

Evaluation on How2Sign. Table 3 shows that SignAlign achieves T2V and V2T R@1 scores of 60.3% and 54.6%, with an $mR@1$ of 57.5%. Relative to UPRet, R@1 improves by 1.2 percentage points in both directions, while MnR decreases from 54.4 to 52.8 for T2V and from 76.4 to 69.7 for V2T. Together, the results across the three benchmarks support the effectiveness of structure-aware alignment for sign language retrieval.

4.3 Ablation Studies

Effect of Different Components. As shown in Table 4, both TSSA and CTSA individually improve $mR@1$ over the baseline on all three datasets. Their combination achieves the highest $mR@1$ scores of 75.0%, 82.3%, and 57.5% on PHOENIX-2014T, CSL-Daily, and How2Sign, respectively.

Table 3: Comparison with baselines on How2Sign. † denotes our RGB-only reproduction of SEDS. Bold indicates the best reported results, including ties.

Method	Text-to-Video					Video-to-Text					mR@1↑
	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	
SA-SR (Duarte et al., 2022)	18.9	32.1	36.5	62.0	–	11.6	27.4	32.5	69.0	–	15.3
SA-CM (Duarte et al., 2022)	24.3	40.7	46.5	16.0	–	17.9	40.1	46.9	14.0	–	21.1
SA-COMB (Duarte et al., 2022)	34.2	48.0	52.6	8.0	–	23.6	47.0	53.0	7.5	–	28.9
CiCo (Cheng et al., 2023)	56.6	69.9	74.7	**1.0**	–	51.6	64.8	70.1	**1.0**	–	54.1
SEDS-RGB† (Jiang et al., 2024a)	54.3	68.8	74.4	**1.0**	70.9	50.3	62.6	68.7	**1.0**	78.8	52.3
UPRet (Wu et al., 2024)	59.1	71.5	75.7	**1.0**	54.4	53.4	65.4	70.0	**1.0**	76.4	56.3
SignAlign (Ours)	**60.3**	**72.6**	**76.6**	**1.0**	**52.8**	**54.6**	**66.6**	**71.5**	**1.0**	**69.7**	**57.5**

Table 4: Effects of TSSA and CTSA on T2V retrieval and bidirectional mR@1.

TSSA	CTSA	PHOENIX-2014T				CSL-Daily				How2Sign			
		R@1↑	R@5↑	R@10↑	mR@1↑	R@1↑	R@5↑	R@10↑	mR@1↑	R@1↑	R@5↑	R@10↑	mR@1↑
×	×	70.7	88.0	93.0	70.9	79.2	90.4	93.6	77.9	55.3	69.7	74.1	52.8
✓	×	73.7	88.6	93.6	74.2	79.8	91.7	94.6	78.9	56.6	70.4	75.3	54.3
×	✓	**74.0**	**89.7**	93.9	73.8	79.6	90.4	93.4	78.3	56.9	70.5	75.0	54.2
✓	✓	**74.0**	89.6	**94.5**	**75.0**	**84.1**	**92.4**	**94.7**	**82.3**	**60.3**	**72.6**	**76.6**	**57.5**

These results exceed the stronger single-component configuration by 0.8, 3.4, and 3.2 percentage points, respectively. The consistent gains support jointly preserving inter-query semantic geometry and coordinating visual evidence allocation within video-text pairs.

Geometric Constraints in TSSA. Table 5 evaluates TSSA constraints on PHOENIX-2014T with CTSA disabled, reporting T2V R@1, V2T R@1, and $mR@1$ in that order. Combining distance and angular constraints achieves the highest $mR@1$ of 74.2%, exceeding the individual variants by 1.3 and 1.8 percentage points, respectively. This supports their complementary contributions to bidirectional retrieval.

Representation Choice for TSSA. Under the same setting, Table 6 compares the representations used for TSSA. EOT-based geometry achieves an $mR@1$ of 74.2%, compared with 72.7% for mean pooling and 70.9% without TSSA. These results favor EOT representations as the semantic anchor for retrieval adaptation.

Evidence Allocation Strategy in CTSA. With TSSA disabled, Table 7 compares native matching, transport-only scoring, and their combination in CTSA on PHOENIX-2014T. Transport alone yields an $mR@1$ of 70.8%, comparable to 70.9% for native matching. Combining the two scores through CTSA improves $mR@1$ to 73.8%, supporting their complementary use for local matching.

Effect of Distance Weight λ_d. Figure 4(a) evaluates the distance constraint on CSL-Daily. Both retrieval directions achieve their highest R@1 at $\lambda_d = 22.5$, followed by a decline at $\lambda_d = 30$. This non-monotonic trend suggests that stronger distance preservation does not necessarily lead to better retrieval. An appropriate weight balances the preservation of pretrained semantic relations with the flexibility required for cross-modal adaptation.

Effect of Angular Weight λ_a. Figure 4(b) shows that nonzero angular weights improve both retrieval directions over $\lambda_a = 0$. T2V R@1 peaks at $\lambda_a = 45$, whereas V2T R@1 peaks at $\lambda_a = 30$, indicating different responses to angular regularization. The best overall $mR@1$ is achieved at $\lambda_a = 45$. Increasing the weight further to $\lambda_a = 60$ reduces performance in both directions, supporting the use of a moderate angular constraint.

Effect of Blend Factor β. Figure 4(c) examines score blending in CTSA on PHOENIX-2014T. Among the tested values, $\beta = 0.075$ achieves the best performance in both directions, outperforming native matching alone ($\beta = 0$) and transport scoring alone ($\beta = 1$). A small transport contribution is

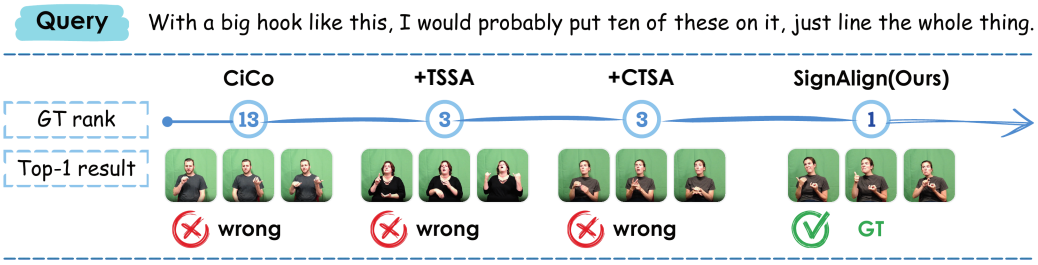


Figure 3: Qualitative comparison of text-to-video retrieval on How2Sign.

Table 5: TSSA constraints.

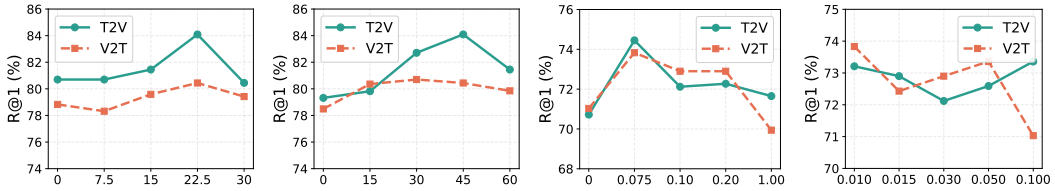
Dist.	Ang.	R@1		mR@1
✓	×	**74.2**	71.5	72.9
×	✓	72.6	72.1	72.4
✓	✓	73.7	**74.7**	**74.2**

Table 6: TSSA representations.

Rep.	R@1		mR@1
w/o TSSA	70.7	71.0	70.9
Mean pool	73.1	72.3	72.7
EOT geom.	**73.7**	**74.7**	**74.2**

Table 7: CTSA strategies.

Strategy	R@1		mR@1
CiCo	70.7	71.0	70.9
Sinkhorn	71.7	69.9	70.8
CTSA	**74.0**	**73.6**	**73.8**



(a) Distance weight λ_d (b) Angular weight λ_a (c) Blend factor β (d) Temperature τ

Figure 4: Sensitivity to TSSA weights on CSL-Daily and CTSA parameters on PHOENIX-2014T.

therefore sufficient to improve retrieval, while assigning it greater weight does not yield further gains over this setting. These results suggest that transport-based evidence allocation is most effective when combined with native matching.

Effect of Transport Temperature τ. Figure 4(d) examines the temperature controlling the smoothness of transport-based evidence allocation. Lower temperatures favor more concentrated transport plans, whereas higher temperatures encourage more diffuse allocations under the same marginal constraints. Among the tested values, $\tau = 0.01$ achieves the highest mR@1. Although $\tau = 0.1$ yields the highest T2V R@1, it substantially reduces V2T R@1, highlighting the need to consider both retrieval directions when selecting the temperature.

4.4 Qualitative Visualization

Figure 3 presents a qualitative text-to-video retrieval example on How2Sign. For the same query, TSSA and CTSA individually improve the GT rank from 13 to 3, while their top-1 predictions remain incorrect. By combining both components, SignAlign retrieves the GT video at rank 1, demonstrating the effectiveness of semantic structure preservation and grounded evidence alignment.

5 Conclusion

In this work, we revisit sign language retrieval by considering semantic relations among texts and joint alignment between video tokens and text tokens. We propose SignAlign to incorporate both structures into cross-modal representation learning. TSSA preserves pretrained distance and angular relations among sentence embeddings during retrieval adaptation. CTSA jointly determines token alignment weights and fuses the resulting similarity score with local aggregation scores, limiting repeated emphasis on a few video tokens. Experiments on multiple public benchmarks demonstrate improved bidirectional retrieval, supporting the effectiveness of structure-aware alignment for sign language retrieval.

REFERENCES

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European conference on computer vision*, pp. 35–53. Springer, 2020.

Sobhan Asasi, Ozge Mercanoglu Sincan, and Richard Bowden. Signseek: Learning transferable representations for sign dictionary retrieval. *arXiv preprint arXiv:2609.03695*, 2026.

Ziyi Bian, Yadan Luo, Fangzhi Zhu, Xiangbo Shu, and Zheng Zhang. Dual-attention video representation learning for parameter efficient text-video retrieval. *IEEE Transactions on Multimedia*, 2026.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793, 2018. doi: 10.1109/CVPR.2018.00812.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, 2020.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in neural information processing systems*, 35:17043–17056, 2022.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

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.

Yasser Hamidullah, Koel Dutta Chowdhury, Yusser Al Ghussin, Shakib Yazdani, Cennet Oguz, Josef Genabith, and Cristina España-Bonet. Grounding or guessing? visual signals for detecting hallucinations in sign language translation. In *International Conference on Learning Representations*, volume 2026, pp. 31868–31891, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

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.

Onur Keleş, Asli Ozyurek, Gerardo Ortega, Kadir Gökgöz, and Esam Ghaleb. The visual iconicity challenge: Evaluating vision-language models on sign language form–meaning mapping. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41101–41116, 2026.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. Leveraging the power of mllms for gloss-free sign language translation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21048–21058. IEEE, 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, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1302. URL https://aclanthology.org/2026.acl-long.1302/.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. In *International Conference on Learning Representations*, volume 2025, pp. 14611–14630, 2025.

Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. In *International Conference on Learning Representations*, pp. 44188–44207, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c0f1eed3f3ca4b2e19ab639a8426270b-Paper-Conference.pdf.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 638–647. Association for Computing Machinery, 2022. doi: 10.1145/3503161.3547910. URL https://doi.org/10.1145/3503161.3547910.

Ariel Basso Madjoukeng, Jérôme Fink, Pierre Poitier, Edith Belise Kenmogne, and Benoit Frenay. Ssl-slr: Self-supervised representation learning for sign language recognition. *arXiv preprint arXiv:2509.05188*, 2025.

WonJun Moon, MinSeok Jung, Gilhan Park, Tae-Young Kim, Cheol-Ho Cho, Woojin Jun, and Jae-Pil Heo. Mitigating semantic collapse in partially relevant video retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Wlpf0Vg4yU.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sihan Tan, Taro Miyazaki, Nabeela Khan, and Kazuhiro Nakadai. Improvement in sign language translation using text ctc alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3255–3266, 2025.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. in 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16852–16861, 2021.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *International Conference on Learning Representations*, volume 2024, pp. 18157–18174, 2024.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Signrep: Enhancing self-supervised sign representations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22804–22814. IEEE, 2025.

Xuan Wu, Hongxiang Li, Yuanjiang Luo, Xuxin Cheng, Xianwei Zhuang, Meng Cao, and Keren Fu. Uncertainty-aware sign language video retrieval with probability distribution modeling. In *European Conference on Computer Vision*, pp. 390–408. Springer, 2024.

Yuanhao Xiong, Long Zhao, Boqing Gong, Ming-Hsuan Yang, Florian Schroff, Ting Liu, Cho-Jui Hsieh, and Liangzhe Yuan. Structured video-language modeling with temporal grouping and spatial grounding. In *International Conference on Learning Representations*, pp. 8911–8928, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25203d1cc8c58381eab578f4fcf9c4f8-Paper-Conference.pdf.

Huijie Yao, Wengang Zhou, Hao Zhou, Hezhen Hu, and Houqiang Li. Retrieval-augmented sign language translation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(2):1–19, 2026.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cpDhcsEDC2.

Shakib Yazdani, Josef Van Genabith, and Cristina España-Bonet. Continual learning in multilingual sign language translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10923–10938, 2025.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. *Advances in Neural Information Processing Systems*, 37:107379–107402, 2024.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7347–7360. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.570. URL https://aclanthology.org/2021.acl-long.570/.

Donglin Zhang, Zheng Rao, and Xiao-Jun Wu. Prototype-based asymmetric hierarchical matching for text-video retrieval. *IEEE Transactions on Multimedia*, 2026a.

Donglin Zhang, Zheng Rao, Xing Xu, Xiao-Jun Wu, and Josef Kittler. Probabilistic embeddings with evidence learning and refinement for text-video retrieval. *IEEE Transactions on Image Processing*, 2026b.

Ruixiang Zhao, Zhihao Xu, Bangxiang Lan, Zijie Xin, Jingyu Liu, and Xirong Li. Save: Speech-aware video representation learning for video-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31166–31175, 2026a.

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, 2026b.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13009–13016, 2020.

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.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23806–23816. IEEE, 2025.

A APPENDIX

USAGE OF LARGE LANGUAGE MODELS

LLM-assisted tools were used to support literature exploration, methodological discussion, code development and debugging, experiment monitoring and analysis, and manuscript preparation. LLM-generated statements were not treated as experimental evidence: numerical results are supported by executable code, recorded configurations, training logs, and checkpoint evaluations. The authors are responsible for validating the results and for the final methodological claims and manuscript.

A.1 ADDITIONAL IMPLEMENTATION DETAILS

Encoders and Input Processing. The video encoder is a Transformer that processes pre-extracted 1024-dimensional RGB features, while the text encoder is initialized from the text tower of CLIP ViT-B/32. Both encoders are optimized during retrieval training, whereas the RGB feature extractor remains fixed. Their outputs are projected into a shared 512-dimensional embedding space. Retrieval training retains CiCo's random word-swap augmentation, while TSSA uses the original, unaugmented texts.

Reference Encoder for TSSA. The reference encoder is copied from the CLIP text encoder at initialization, including its pretrained projection layer. Its parameters remain fixed throughout retrieval training. It is kept in evaluation mode, and its outputs are computed without gradients. For TSSA, the trainable and reference encoders process the same original, unaugmented texts. We extract the projected EOT embedding of each sentence from both encoders and compare the pairwise distances and triplet-wise angles within the mini-batch. Distances from each encoder are divided by that encoder's own mean nonzero pairwise distance. The reference encoder is discarded at inference.

Token Selection for CTSA. CTSA constructs its transport matrix using only valid video-segment and text-content tokens. On the video side, padding positions and the classification token are excluded. On the text side, padding positions and the start-of-text and EOT tokens are excluded. Thus, F and L denote the numbers of retained visual and textual tokens for each candidate pair, and the uniform transport marginals are set to $1/F$ and $1/L$, respectively. These exclusions apply only to transport computation. The native matching branches retain CiCo's original token selection and aggregation rules (Cheng et al., 2023).

Transport Computation. We approximate the transport plan using Sinkhorn iterations in the log domain. Given the token–segment similarity matrix M, we define $A = M/\tau$, where τ is the transport temperature. The log scaling factors α_f and η_ℓ adjust the weights associated with visual segment f and text token ℓ, respectively. Starting from zero, they are updated alternately:

$$\begin{aligned} \alpha_f &\leftarrow -\log F - \mathrm{LSE}_\ell \left(A_{f\ell} + \eta_\ell\right), \\ \eta_\ell &\leftarrow -\log L - \mathrm{LSE}_f \left(A_{f\ell} + \alpha_f\right), \end{aligned} \tag{10}$$

where $\mathrm{LSE}_j(x_j) = \log \sum_j \exp(x_j)$. The first update normalizes each visual row to a total weight of $1/F$. The second uses the updated row factors to normalize each textual column to $1/L$. Because column normalization can change the row sums, the two updates are repeated to approximate both marginal constraints. After K iterations, we recover the weights and normalize their total mass:

$$Q_{f\ell} = \exp\left(A_{f\ell} + \alpha_f + \eta_\ell\right), \qquad \widehat{P}_{f\ell} = \frac{Q_{f\ell}}{\sum_{f'=1}^{F} \sum_{\ell'=1}^{L} Q_{f'\ell'}}. \tag{11}$$

The resulting matrix $\widehat{P}$ is used to compute the transport score defined in the main text. Log-domain updates improve numerical stability, while the finite iteration budget may leave residual deviations from the prescribed marginals. We use $K = 10$ for the main-text PHOENIX-2014T configuration. Gradients propagate through these operations during training, and the same transport computation is retained at inference.

Computational Cost. CTSA introduces no additional trainable parameters. For a pair with F visual tokens, L text tokens, and embedding dimension d, similarity computation requires $O(FLd)$

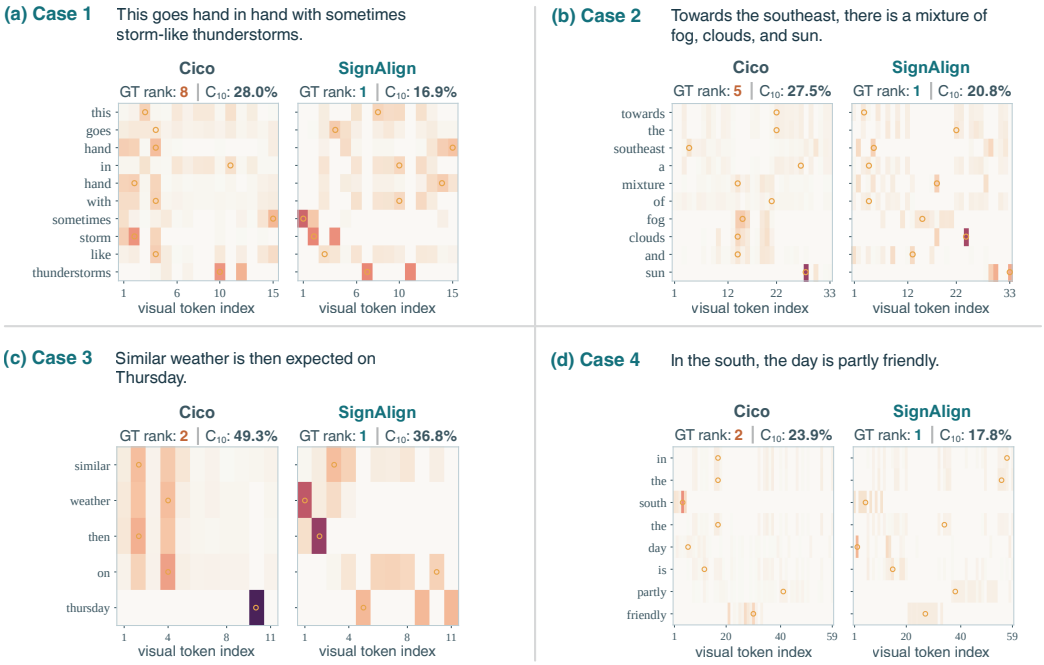


Figure 5: Local evidence allocation for CiCo and SignAlign on four PHOENIX-2014T examples. Rows correspond to word positions and columns to visual tokens. Circles mark the highest-weight position in each row. GT rank denotes the ground-truth video's rank in text-to-video retrieval. C_{10} measures the proportion of matching weight assigned to the most-used 10% of visual tokens.

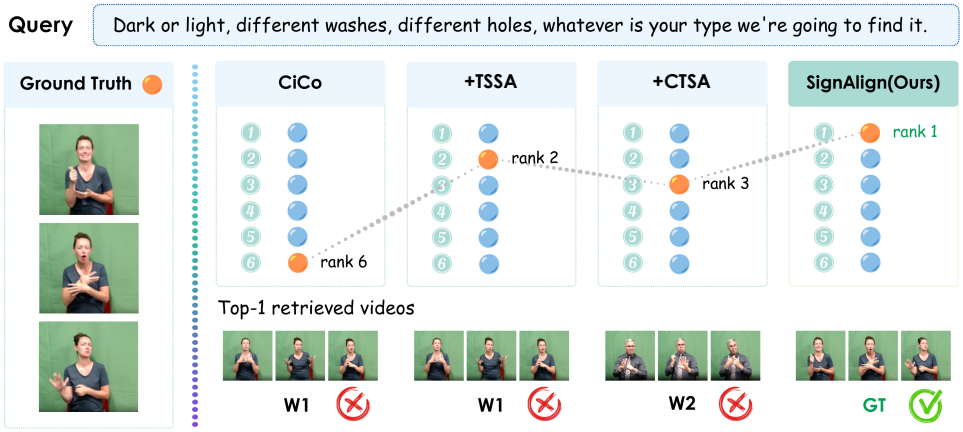


Figure 6: Text-to-video retrieval comparison on How2Sign. Orange markers indicate the ground-truth video's rank. Frames below each ranked list show the top-1 retrieved video. W1 and W2 denote incorrect videos, while GT denotes the ground-truth video.

operations. Sinkhorn updates and transport scoring require an additional $O(KFL)$ and $O(FL)$ operations, respectively. The similarity and transport matrices each occupy $O(FL)$ storage, with intermediate states also retained for differentiation during training. Each candidate pair is scored independently, so gallery token features can be cached.

Table 8: Effect of marginal constraints on T2V retrieval and bidirectional mR@1 (%). TSSA is disabled, with $\beta = 0.1$ and $\tau = 0.03$. K denotes the number of Sinkhorn iterations; – denotes non-iterative normalization. Bold indicates column-wise best results, including ties.

Constraints	K	PHOENIX-2014T				CSL-Daily			
		R@1↑	R@5↑	R@10↑	mR@1↑	R@1↑	R@5↑	R@10↑	mR@1↑
Global normalization	–	69.94	88.63	92.83	70.64	78.82	91.10	**93.73**	78.19
Row only	–	70.72	**90.97**	**95.02**	70.64	79.95	89.97	92.36	79.01
Column only	–	72.27	89.41	93.30	71.50	80.58	**91.35**	**93.73**	79.11
Row and column	3	72.12	89.41	93.46	72.51	78.20	89.22	92.23	77.66
Row and column	5	**73.36**	89.88	94.39	**73.05**	**80.95**	90.73	92.73	**79.85**

Table 9: Effect of evaluation scoring on PHOENIX-2014T (%). The last two rows use identical encoder weights.

Training	Evaluation score	T2V R@1	V2T R@1	mR@1
Baseline	Native	70.72	71.03	70.87
Baseline + CTSA	Native	73.68	73.68	73.68
Baseline + CTSA	Native + transport	**74.45**	**73.83**	**74.14**

A.2 Additional Experimental Results

Experimental Settings. The marginal-constraint and Sinkhorn-iteration experiments disable TSSA and fix $\beta = 0.1$ and $\tau = 0.03$. Each configuration is trained for 200 epochs with batch size 128 and random seed 42. We use BertAdam with a learning rate of 10^{-5} for both retrieval encoders, weight decay 10^{-3}, a warm-up ratio of 0.1, and cosine learning rate scheduling. The maximum text and video sequence lengths are 32 and 64, respectively. Random word-swap augmentation is applied with probability 0.5.

Effect of Marginal Constraints. We examine whether constraining both visual and textual matching weights improves retrieval over global or one-sided normalization. Rows of M correspond to video segments, and columns correspond to text tokens. Global normalization applies a single softmax over all entries of M/τ. Row-only normalization applies softmax within each row and scales the weights by $1/F$, fixing each segment's total weight. Column-only normalization applies softmax within each column and scales the weights by $1/L$, fixing each text token's total weight. Joint normalization imposes both constraints through Sinkhorn iterations. All variants produce unit-mass matching matrices and use the same weighted-similarity score and blending coefficient. The global-normalization variant is a control for transport scoring, not the native CiCo scorer.

Table 8 compares these variants with joint normalization at $K = 3$ and $K = 5$. At $K = 5$, joint constraints achieve the highest T2V R@1 and bidirectional $mR@1$ on both datasets. Compared with the stronger one-sided variant, $mR@1$ increases by 1.55 percentage points on PHOENIX-2014T and 0.74 points on CSL-Daily. This advantage is not observed on CSL-Daily at $K = 3$, where joint normalization reaches 77.66%, below both one-sided variants. The improvement is also concentrated in top-1 retrieval: row-only normalization gives the highest R@5 and R@10 on PHOENIX-2014T, while column-only normalization gives the highest R@5 and a tied highest R@10 on CSL-Daily.

Effect of Inference Scoring. We examine whether transport scoring improves retrieval when the encoder representations remain unchanged. For this comparison, we use a CTSA model trained with TSSA disabled and $(\beta, \tau, K) = (0.075, 0.03, 3)$ on PHOENIX-2014T. We evaluate its epoch 151 checkpoint twice: first using only the native matching scores, then blending each native score with the transport score. No parameters are updated between evaluations. As shown in Table 9, native scoring with these encoders achieves 73.68% mR@1, compared with 70.87% for the baseline checkpoint at epoch 200. Adding the transport score to the same encoder outputs increases T2V R@1 from 73.68% to 74.45% and V2T R@1 from 73.68% to 73.83%, reaching 74.14% mR@1.

Table 10: Effect of Sinkhorn iterations on T2V retrieval and bidirectional mR@1 (%). TSSA is disabled, with $\beta = 0.1$ and $\tau = 0.03$.

Iterations K	PHOENIX-2014T				CSL-Daily			
	R@1↑	R@5↑	R@10↑	mR@1↑	R@1↑	R@5↑	R@10↑	mR@1↑
1	72.74	89.72	94.24	72.43	78.32	**90.98**	**93.48**	77.81
3	72.12	89.41	93.46	72.51	78.20	89.22	92.23	77.66
5	73.36	**89.88**	94.39	73.05	**80.95**	90.73	92.73	**79.85**
10	**74.00**	89.70	93.90	**73.80**	80.33	90.23	93.11	79.53
20	73.05	89.72	**94.70**	72.66	80.20	90.10	93.36	79.68

Table 11: Marginal errors on five fixed PHOENIX-2014T pairs at $\tau = 0.03$. Lower values indicate closer agreement with the prescribed marginals.

K	Mean ϵ_{vis}	Max ϵ_{vis}	Max ϵ_{text}
1	0.5181	0.6707	1.40×10^{-7}
3	0.2825	0.4016	2.69×10^{-7}
5	0.1727	0.2598	2.30×10^{-7}
10	0.0818	0.1389	2.25×10^{-7}
20	0.0286	0.0629	2.12×10^{-7}

Among the 642 text queries, this scoring change corrects six previously incorrect top predictions and makes one previously correct prediction incorrect. The comparison between the last two rows therefore shows an additional retrieval gain from transport scoring without further encoder training.

Effect of Sinkhorn Iterations. We vary the number of Sinkhorn iterations K while disabling TSSA and fixing $\beta = 0.1$ and $\tau = 0.03$. Each configuration is trained separately with the same K used during training and inference. As shown in Table 10, PHOENIX-2014T achieves its highest mR@1 of 73.80% at $K = 10$, whereas CSL-Daily reaches 79.85% at $K = 5$. Increasing K to 20 reduces mR@1 from these peaks by 1.14 and 0.17 percentage points, respectively. The best R@5 and R@10 occur at different iteration counts, so no single setting achieves the highest recall at every rank. These results support selecting the iteration count by retrieval performance rather than simply increasing the number of normalization steps.

Accuracy of Marginal Constraints. We evaluate how closely the computed transport weights satisfy the prescribed marginals. Using fixed similarity matrices from five selected ground-truth video–text pairs on PHOENIX-2014T, we vary the Sinkhorn iteration count K at $\tau = 0.03$ without updating the encoders. For each pair, the visual and textual marginal errors are

$$\epsilon_{\text{vis}} = \sum_{f=1}^{F} \left| \sum_{\ell=1}^{L} \widehat{P}_{f\ell} - \frac{1}{F} \right|, \qquad \epsilon_{\text{text}} = \sum_{\ell=1}^{L} \left| \sum_{f=1}^{F} \widehat{P}_{f\ell} - \frac{1}{L} \right|. \tag{12}$$

Here, ϵ_{vis} sums the absolute differences between the visual row sums and their target $1/F$, while ϵ_{text} measures the corresponding column errors relative to $1/L$. Transport weights are computed in FP32, and errors are accumulated in FP64. Table 11 reports the mean and maximum errors across the five pairs. The mean visual error decreases from 0.5181 at $K = 1$ to 0.0818 at $K = 10$ and 0.0286 at $K = 20$. The textual error remains on the order of 10^{-7} because each iteration ends with column normalization. Thus, additional iterations improve the visual marginal accuracy on these examples, while a nonzero error remains even at $K = 20$. Unlike the separately trained models in Table 10, this comparison measures only the numerical effect of additional iterations on unchanged similarities.

A.3 Additional Qualitative Analysis

Local Evidence Allocation. Figure 5 compares the local matching patterns of CiCo and SignAlign on four PHOENIX-2014T examples. Rows represent text word positions, and columns represent visual tokens. Darker colors indicate larger matching weights, and circles mark the highest-weight visual position for each word.

We quantify evidence concentration using C_{10}, the proportion of total matching weight assigned to the most-used 10% of visual tokens. Specifically, let $p(f \mid w)$ denote the conditional matching weight of visual token f for word position w. For W word positions and F visual tokens, we define

$$u_f = \frac{1}{W} \sum_{w=1}^{W} p(f \mid w), \qquad C_{10} = \sum_{f \in \mathcal{I}_{10}} u_f, \tag{13}$$

where u_f is the average weight received by visual token f, and $\mathcal{I}_{10}$ contains the indices of the $\lceil 0.1F \rceil$ largest values of u_f. Lower C_{10} indicates less concentrated evidence allocation.

Across the four examples, SignAlign reduces C_{10} from 28.0% to 16.9%, 27.5% to 20.8%, 49.3% to 36.8%, and 23.9% to 17.8%, respectively. In Case 3, for example, CiCo assigns the highest weights for multiple words to the same visual positions, whereas SignAlign distributes these maxima across different positions. The ground-truth retrieval ranks also improve from 8, 5, 2, and 2 to 1 in all four cases. These examples show less concentrated matching alongside improved retrieval, although lower concentration alone does not establish more accurate word–sign correspondence.

Retrieval Comparison. Figure 6 presents a text-to-video retrieval example on How2Sign, showing both the ground-truth video's rank and the top-1 prediction under each model configuration. For the query "Dark or light, different washes, different holes, whatever is your type we're going to find it," CiCo ranks the ground-truth video sixth and retrieves an incorrect video, W1, at rank one. Adding TSSA raises the ground-truth video to second place, although W1 remains the top prediction. Adding CTSA instead ranks the ground-truth video third and changes the top prediction to another incorrect video, W2. Thus, both components improve the position of the correct video, but neither alone places it above all competing candidates.

With both components enabled, SignAlign ranks the ground-truth video first. The frames below the ranked lists show the corresponding change from incorrect top-1 predictions to the correct video. This example illustrates how preserving sentence-level semantic relations and coordinating token-to-segment matching can complement each other: their combination resolves a retrieval error that remains under either component alone.