MULTI-GRANULAR SEMANTIC MATCHING WITH UNCERTAINTY-AWARE CALIBRATION FOR SIGN LANGUAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign language retrieval requires precise cross-modal alignment between sign and textual semantics. However, existing methods typically rely on either global or fine-grained alignment between text and sign video, making it difficult to jointly capture overall event semantics and subtle motion details. Moreover, given a text (respectively a sign video), there are multiple semantically consistent sign videos (respectively texts), and the same sign articulation may convey distinct semantics across different contexts. This correspondence multiplicity and contextual ambiguity introduce uncertainty into sign-text matching, leading to suboptimal semantic alignment. To address these challenges, we propose MGUC, a multi-granular semantic matching framework with uncertainty-aware calibration for sign language retrieval. Specifically, coarse-grained global matching captures overall event semantics and ensures holistic sign–text consistency. Fine-grained local matching captures subtle motion details by jointly modeling cross-modal relevance and local feature interactions through bidirectional similarity aggregation, with learnable `[Null]` tokens mitigating spurious matches arising from transitional clips and words without explicit sign gestures. We further introduce an uncertainty-aware modeling and calibration module to capture uncertainty in sign–text correspondences and refine cross-modal alignment. It represents sign videos and texts as multivariate Gaussian distributions and uses Bhattacharyya distances to quantify their distributional discrepancy. The relative distance between unpaired and paired samples is used to adaptively calibrate the negative contributions during contrastive learning. This reduces excessive repulsion from potentially semantically related samples, thereby refining text-to-sign alignment in the shared embedding space. Extensive experiments on three widely used benchmarks demonstrate that MGUC achieves competitive retrieval performance.

1 INTRODUCTION

Sign language is a primary means of communication for many deaf and hard-of-hearing people. Sign language understanding is an important research field dedicated to overcoming communication barriers between deaf-mute and non-signers. Research on sign-language understanding has largely focused on recognition (Guo et al., 2023; Albanie et al., 2020; Pu et al., 2019; Cui et al., 2019; Camgoz et al., 2020), which predicts gloss sequences from signing videos, and translation (Voskou et al., 2021; Zhou et al., 2021b;a), which maps signing videos to natural-language text. In this paper, we focus on the emerging sign language retrieval task (Cheng et al., 2023; Jiang et al., 2024a; Wu et al., 2024). Unlike sign language recognition and translation, the sign language retrieval task aims to retrieve the sign videos or texts most relevant to a query.

Existing retrieval methods (Duarte et al., 2022; Jiang et al., 2024b; Cheng et al., 2023) typically align global representations of sign videos and text queries to capture overall semantic correspondence. However, global matching alone may overlook subtle motion details, which are difficult to capture with a single global embedding. As shown in Fig. 1(a), the first sample shares the overall action "*wash your face*" with the query and may therefore receive a high similarity score under global matching alone. However, it lacks the action detail "*with a mild cleanser*," which is crucial for

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.



(a) Challenge one: Insufficient Semantic Matching **(b) Challenge two: Uncertainty in Sign–Text Matching**

Figure 1: Two challenges in sign-language retrieval. (a) Insufficient semantic matching: global matching may fail to adequately capture fine-grained details (Sample 01), while existing fine-grained matching can be dominated by some strong clip-word partial matches (Sample 02). (b) Uncertainty in sign–text matching: an unpaired sample may depict an event that is semantically consistent with the text query (Sample 01), while the same sign articulation may convey a distinct meaning in different semantic contexts (Sample 02).

precise retrieval. To capture this fine-grained correspondence, recent works (Cheng et al., 2023; Jiang et al., 2024a) construct clip–word similarity matrices and aggregate them using mean-max pooling. However, a few strong local matches may dominate the resulting score without ensuring consistency in the overall semantics. For example, the second sample in Fig. 1(a) shares the locally matched actions "*wash*" and "*dry*" with the query but describes a different event. Consequently, these isolated matches may dominate the aggregated similarity under the mean-max paradigm even when the overall semantics do not match.

Furthermore, a sign video may correspond to multiple semantically related texts, and the same sign articulation may convey distinct semantics in different contexts. These make the semantic relationships between sign videos and texts uncertain. As illustrated in Fig. 1(b), we show a text query, its paired sign video, and two unpaired sign-video samples with their text descriptions. The first sample depicts an event semantically consistent with the query (i.e., "*putting ice into a glass*") even though it is not treated as a matched sample for the given query. In the paired sign video (Ground Truth) and the second sample, the green boxes highlight the same sign articulations that convey distinct meanings in different semantic contexts: this articulation in the paired video corresponds to "*ice*", whereas that in the second sample conveys "*freeze*". Existing methods (Cheng et al., 2023; Jiang et al., 2024a; Wu et al., 2024) leverage vanilla contrastive alignment or optimal-transport based alignment to model text-sign correspondences. Nevertheless, they treat unpaired samples as negatives regardless of their semantic relevance to the query. This may over-penalize semantically relevant samples and distort the text-sign alignment in the shared embedding space.

To address these challenges, we propose a multi-granular semantic matching framework with uncertainty-aware calibration for sign language retrieval, as shown in Fig. 2. To establish comprehensive semantic matching, we introduce coarse-grained global semantic matching (GSM) to enable holistic consistency between sign videos and text. We also introduce fine-grained local semantic matching (FSM), which jointly models cross-modal relevance and local feature interactions through bidirectional similarity aggregation to capture subtle motion details. Within fine-grained matching, we incorporate learnable `[Null]` tokens into both modalities to mitigate spurious correspondences arising from transitional clips and words without explicit sign gestures. To model uncertainty in sign–text matching and refine cross-modal alignment, we propose an uncertainty-aware modeling and calibration module. It represents sign videos and text descriptions as multivariate Gaussian distributions and quantifies the distributional discrepancy of each sign–text pair using the Bhattacharyya distance. We then leverage these distances to evaluate each unpaired pair relative to its corresponding paired sample and adaptively calibrate its contribution as a negative during con-

trastive learning. This reduces excessive repulsion from potentially semantically related samples, thereby refining text-to-sign alignment in the shared embedding space.

Our main contributions are summarized as follows:

- We propose MGUC, a multi-granular semantic matching framework with uncertainty-aware calibration for sign-language retrieval. This framework conducts cross-modal semantic alignment across multiple granularities and explicitly captures the inherent uncertainty of sign–text matching.
- We develop multi-granular semantic matching for comprehensive sign–text alignment. Coarse-grained global matching captures overall event semantics and ensures holistic consistency. Fine-grained local matching captures motion details through interaction-aware bidirectional similarity aggregation, with learnable `[Null]` tokens mitigating spurious matches from transitional clips and words without explicit sign gestures.
- We introduce uncertainty-aware modeling and calibration by representing sign videos and text descriptions as multivariate Gaussian distributions. We leverage Bhattacharyya distances to evaluate unpaired samples relative to their corresponding paired samples and adaptively calibrate their negative contributions during contrastive learning.
- Extensive experiments and comprehensive visualization analyses on three widely used sign language benchmarks, *How2Sign*, *PHOENIX-2014T*, and *CSL-Daily*, demonstrate the effectiveness of our proposed method.

2 Related Work

Sign Language Retrieval. Sign-language retrieval (SLRet) aims to retrieve relevant sign-video samples given text queries. Existing methods typically encode a temporally ordered signing sequence and a sentence into a shared space, then optimize bidirectional ranking with paired video–text examples. CiCo (Cheng et al., 2023) proposes cross-lingual contrastive learning and introduces local video–text correspondence for sign-language retrieval. UPRet (Wu et al., 2024) employs optimal transport to establish fine-grained alignment between sign videos and text. SEDS (Jiang et al., 2024a) improves sign representations through complementary pose and RGB streams, combining local action details with global visual context. Compared with these methods, we propose a unified multi-granular matching framework that captures both holistic semantics and fine-grained motion details, with an uncertainty-aware calibration strategy that reweights the negative contributions of unpaired samples during contrastive learning.

Cross-Modal Semantic Alignment. Cross-modal semantic alignment (Ma et al., 2022; Li et al., 2022; Wang et al., 2021; Dong et al., 2023; Yang et al., 2024; Zhao et al., 2023; Lee et al., 2018) is a significant task in multi-modality scenarios, which aims to bring together instances from different modalities that share similar semantics. FSM (Zhu et al., 2023) introduces the FIMA framework, which combines dense contrastive learning, foreground sampling, and a motion decoder to address weak alignment between modalities at the pixel and frame levels. CASR (Chen et al., 2026) uses missing-modality imagination to recover semantic information in incomplete cross-modal retrieval. It further employs explicit causal alignment to suppress spurious cross-modal correlations. DecAlign (Qian et al., 2026) proposes a hierarchical cross-modal alignment framework that decouples modality-common and modality-specific representations using prototype-guided optimal transport and distribution-level alignment. Due to the subtle motion details and context-dependent meanings of signs, directly applying these alignment strategies to sign language retrieval remains challenging.

Probabilistic Representations. Probabilistic modeling has been widely studied in machine learning. HIB (Oh et al., 2018) introduced probabilistic embeddings to capture image-feature uncertainty, with subsequent applications to face recognition (Shi & Jain, 2019; Chang et al., 2020) and pose estimation (Sun et al., 2020). For cross-modal retrieval, PCME (Chun et al., 2021) models one-to-many text–image correspondences with probabilistic embeddings, while some works (Fang et al., 2023; Wu et al., 2024) employ multiple-instance learning via Monte Carlo sampling and optimal transport for fine-grained alignment. By contrast, we compute Bhattacharyya distances in closed form from Gaussian parameters without sampling. We then use these distances relative to paired samples to calibrate the contributions of unpaired pairs during contrastive learning. This calibration is integrated with multi-granular matching to refine sign–text alignment.

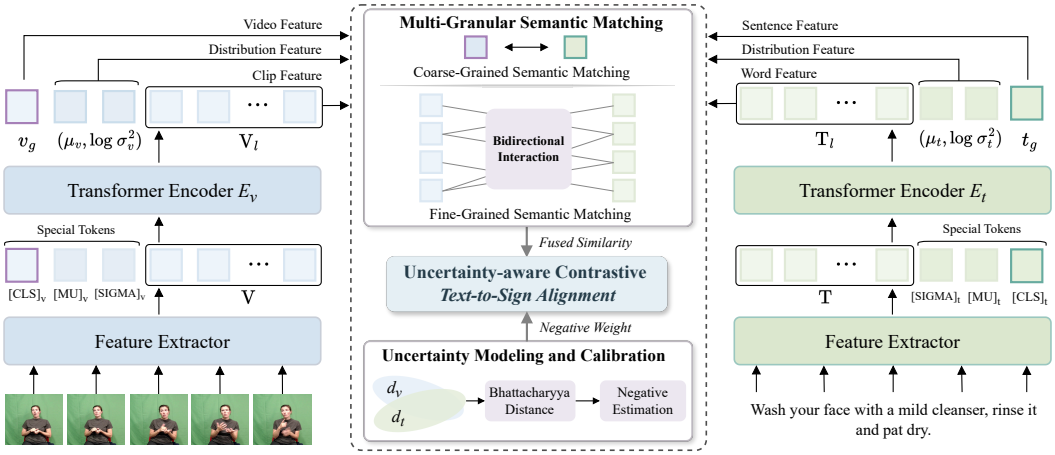

Figure 2: Overview of MGUC. We encode sign videos and text queries into global, local, and distributional representations. The multi-granular semantic matching module combines coarse-grained global matching (GSM) and fine-grained local matching (FSM) to produce a fused sign–text similarity. The uncertainty-aware modeling and calibration module uses Bhattacharyya distances relative to paired samples to compute weights for unpaired pairs. These weights calibrate their negative contributions to contrastive learning based on the fused similarity.

3 METHOD

This section presents MGUC in detail, as shown in Fig. 2. Given a sign video and a text query, we first extract global, local, and distributional representations for both modalities (Sec. 3.1). We then combine coarse-grained global semantic matching with fine-grained local semantic matching to capture overall semantics and subtle motion details. Their similarities are fused into a sign–text retrieval score (Sec. 3.2). Finally, we measure the discrepancy between the distributional representations using the Bhattacharyya distance and calibrate the contributions of unpaired pairs during contrastive training (Sec. 3.3).

3.1 SIGN AND TEXT REPRESENTATIONS

Given a sign language video $\mathcal{V}$, we first divide it into temporally ordered clips following SEDS (Jiang et al., 2024a). We then employ an I3D network pre-trained on BSL-1K (Albanie et al., 2020) as the offline visual encoder to extract clip-level sign features:

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{N_v}] \in \mathbb{R}^{N_v \times D}, \tag{1}$$

where N_v denotes the number of sampled sign clips, and D is the feature dimension.

For a natural language query $\mathcal{T}$ containing N_t words, we first adopt a pre-trained frozen DistilBERT (Sanh et al., 2019) to extract the word representations:

$$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{N_t}] \in \mathbb{R}^{N_t \times D}. \tag{2}$$

Then, we build Transformer encoders $\mathcal{E}_v$ and $\mathcal{E}_t$, which take clip features and word features with additional learnable tokens as inputs to obtain global embeddings, probabilistic embeddings, and local embeddings. Specifically, inspired by the class token in BERT (Devlin et al., 2019) and ViT (Dosovitskiy et al., 2021), we prepend [CLS], [MU], and [SIGMA] tokens to both features:

$$[\mathbf{v}_g, \boldsymbol{\mu}_v, \log \boldsymbol{\sigma}_v^2, \mathbf{V}_l] = E_v([[\mathrm{CLS}]_v, [\mathrm{MU}]_v, [\mathrm{SIGMA}]_v, \mathbf{V}]), \tag{3}$$

$$[\mathbf{t}_g, \boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t^2, \mathbf{T}_l] = E_t([[\mathrm{CLS}]_t, [\mathrm{MU}]_t, [\mathrm{SIGMA}]_t, \mathbf{W}]), \tag{4}$$

where E_v and E_t are the Transformer encoders, $\mathbf{v}_g$ and $\mathbf{t}_g$ denote the global embeddings. $\boldsymbol{\mu}_v$, $\log \boldsymbol{\sigma}_v^2$, $\boldsymbol{\mu}_t$, and $\log \boldsymbol{\sigma}_t^2$ are the outputs corresponding to the [MU] and [SIGMA] tokens. $\mathbf{V}_l$ and $\mathbf{T}_l$ denote the clip-level and word-level features. Within each Transformer encoder, the distributional tokens interact with both global and local tokens to model uncertainty informed by global context and local details.

3.2 Multi-Granular Semantic Matching

3.2.1 Global Semantic Matching

The global branch captures the overall semantic correspondence between the complete textual query and sign video. After text-sign representation learning, we obtain the global embeddings $\mathbf{t}_g$ and $\mathbf{v}_g$. For the m-th text and the n-th sign video, their global similarity is computed by cosine similarity:

$$\mathbf{S}_{\text{glo}}(m,n) = \mathbf{t}_g^m (\mathbf{v}_g^n)^\top. \tag{5}$$

The resulting $\mathbf{S}_{\text{glo}}$ captures the global semantic correspondence between sign videos and textual descriptions, enabling holistic cross-modal consistency.

3.2.2 Fine-Grained Semantic Matching

Existing fine-grained methods (Cheng et al., 2023; Jiang et al., 2024a) typically aggregate clip–word similarities using mean–max pooling, which can overemphasize a few strong partial matches. Moreover, transitional clips and words without explicit sign gestures may lead to spurious local correspondences. To address these issues, we introduce an interaction-aware aggregation strategy with learnable `[Null]` tokens. This strategy jointly considers cross-modal relevance and local feature interactions, while allowing clips and words without clear matches to align with the corresponding `[Null]` tokens, as shown in Fig. 3. Specifically, given the sign clip feature $\mathbf{V}_l = [\mathbf{v}_1, \ldots, \mathbf{v}_{N_v}]$ and the contextualized word features $\mathbf{T}_l = [\mathbf{w}_1, \ldots, \mathbf{w}_{N_t}]$, we append a learnable `[Null]` $\in \mathbb{R}^D$ token to each modality. For the m-th text and the n-th sign video, we compute the pairwise cosine similarities as

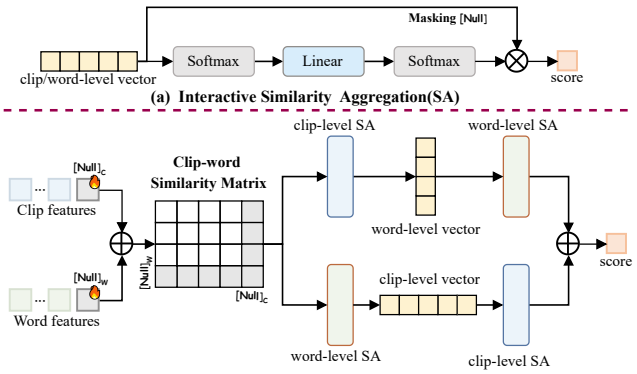

Figure 3: (a) Similarity aggregation learns weights from a clip- or word-level similarity vector, with the `[Null]` entry masked before weighted summation. (b) Bidirectional aggregation processes the clip–word similarity matrix augmented with `[Null]` tokens in clip-to-word and word-to-clip directions. The two resulting scores are averaged to obtain the fine-grained sign–text similarity.

$$\mathbf{C}_{mn}(j,i) = \mathbf{v}_j^n \mathbf{w}_i^{m\top}, \quad \mathbf{C}_{mn} \in \mathbb{R}^{(N_v+1)\times(N_t+1)}, \tag{6}$$

where the last row and column correspond to the visual and textual `[Null]` tokens, respectively. These tokens absorb transitional gestures and words without explicit sign gestures. Their similarities participate in aggregation computation but are excluded from the final weighted summation.

Bidirectional Interaction-Aware Similarity Aggregation (Bi-SA). Although the similarity matrix provides fine-grained cross-modal correspondences, directly applying the commonly used mean-max pooling does not explicitly model interactions among local features during aggregation. We first apply softmax to the input similarity vector to obtain cross-modal relevance weights. A learnable linear transformation then models local feature interactions through these weights. A second softmax produces the final aggregation weights:

$$\mathrm{SA}(\mathbf{x}) = \sum_{k=1}^{K} \left[\mathrm{Softmax}\left(\mathbf{W}\,\mathrm{Softmax}(\mathbf{x}) + \mathbf{b}\right)\right]_k x_k. \tag{7}$$

Here, $\mathbf{x} = [x_1, \ldots, x_K, x_{\text{null}}]$ contains both valid similarities and the `[Null]` similarity. We apply SA bidirectionally to aggregate the clip–word similarity matrix. Denoting the clip-level and word-level aggregation operators as SA_v and SA_t, respectively, we first apply SA_v to obtain a word-level similarity vector, which is then aggregated by SA_t. The final fine-grained similarity is computed as

$$\mathbf{S}_{\text{loc}}(m,n) = \frac{1}{2}\left[\mathrm{SA}_t\left(\mathrm{SA}_v(\mathbf{C}_{mn})\right) + \mathrm{SA}_v\left(\mathrm{SA}_t(\mathbf{C}_{mn})\right)\right]. \tag{8}$$

3.2.3 MULTI-GRANULAR FUSION

The global and fine-grained branches capture complementary aspects of sign–text correspondence. Instead of directly summing their similarities, we introduce learnable normalized coefficients to balance their relative contributions and combine them into a unified retrieval score:

$$\mathbf{S}(m,n) = a_{\text{glo}}\mathbf{S}_{\text{glo}}(m,n) + a_{\text{loc}}\mathbf{S}_{\text{loc}}(m,n), \quad a_k = \frac{\exp(\eta_k)}{\sum_{j\in\{\text{glo,loc}\}}\exp(\eta_j)}, \tag{9}$$

where η_{glo} and η_{loc} are learnable parameters. The normalized scales balance the contributions of the two branches while preserving their complementary information.

3.3 UNCERTAINTY-AWARE MODELING AND CALIBRATION

To model uncertainty in sign–text matching, we represent sign videos and text queries as multivariate Gaussian distributions and compute the Bhattacharyya distance between each video–text pair. We then leverage Bhattacharyya distances to evaluate unpaired samples relative to their corresponding paired samples and adaptively calibrate their negative contributions during contrastive learning.

Probabilistic Distribution Modeling (PDM). We represent each sign video and text query as a multivariate Gaussian distribution. The encoders predict mean and log-variance vectors, from which we construct the diagonal covariance matrices as $\mathbf{\Sigma}_v = \text{diag}(\exp(\log \boldsymbol{\sigma}_v^2))$ and $\mathbf{\Sigma}_t = \text{diag}(\exp(\log \boldsymbol{\sigma}_t^2))$. For the sign video and text query, the corresponding distributions are

$$\mathbf{d}_t \sim \mathcal{N}(\boldsymbol{\mu}_t; \mathbf{\Sigma}_t), \mathbf{d}_v \sim \mathcal{N}(\boldsymbol{\mu}_v; \mathbf{\Sigma}_v) \tag{10}$$

Unlike existing methods that rely on an additional uncertainty head, our distributional representations are learned jointly with global and local semantics. The distribution parameters model uncertainty in cross-modal matching and provide auxiliary supervision for refining the shared embedding space. To prevent the learned probabilistic embedding from collapsing, we employ the KL divergence between the learned probabilistic distributions and the unit Gaussian prior $\mathcal{N}(\mathbf{0};\mathbf{I})$:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\mathbf{d}_t \,\|\, \mathcal{N}(\mathbf{0};\mathbf{I})) + D_{\text{KL}}(\mathbf{d}_v \,\|\, \mathcal{N}(\mathbf{0};\mathbf{I}))\,, \tag{11}$$

Existing distributional approaches often use multiple-instance learning (MIL) (Dietterich et al., 1997) or match probability (MP) (Chun et al., 2021; Fang et al., 2023) to measure the distance of sampled embeddings to approximate the distribution discrepancy. However, a finite set of sampled embeddings may not adequately model the full distribution. Instead, we quantify the distributional discrepancy between Gaussian representations by computing the Bhattacharyya distance directly from their parameters:

$$d(m,n) = \frac{1}{8}(\boldsymbol{\mu}_t^m - \boldsymbol{\mu}_v^n)^\top (\mathbf{\Sigma}^{mn})^{-1}(\boldsymbol{\mu}_t^m - \boldsymbol{\mu}_v^n) + \frac{1}{2}\ln\frac{|\mathbf{\Sigma}^{mn}|}{\sqrt{|\mathbf{\Sigma}_t^m|\,|\mathbf{\Sigma}_v^n|}} \tag{12}$$

Here, $|\cdot|$ denotes the determinant, $\mathbf{\Sigma}^{mn} = (\mathbf{\Sigma}_t^m + \mathbf{\Sigma}_v^n)/2$. A smaller Bhattacharyya distance indicates that the two distributions are closer. We use the $-d(m,n)$ as matching scores and optimize the Gaussian representations with a bidirectional cross-entropy loss, denoted by $\mathcal{L}_{\text{PDM}}$.

Uncertainty-Aware Contrastive Calibration (UACC). We compute a negative weight for each unpaired pair based on its Bhattacharyya distance relative to that of the corresponding paired sample. Specifically, for text-to-video and video-to-text retrieval, respectively, we define:

$$\Delta_{mn}^{t2v} = d(m,n) - d(m,m), \quad \Delta_{nm}^{v2t} = d(n,m) - d(m,m). \tag{13}$$

A small distance gap indicates that the unpaired pair has a Bhattacharyya distance close to that of the paired pair, whereas a large positive gap indicates a substantially larger cross-modal discrepancy. We convert these gaps into negative weights:

$$\omega_{mn}^{t2v} = 1 - \gamma\,\text{sg}\big[\text{sigmoid}\big(-\Delta_{mn}^{t2v}/\tau_a\big)\big]\,, \quad \omega_{nm}^{v2t} = 1 - \gamma\,\text{sg}\big[\text{sigmoid}\big(-\Delta_{nm}^{v2t}/\tau_a\big)\big]\,, \tag{14}$$

where τ_a is a temperature, γ sets the maximum down-weighting, and $\text{sg}(\cdot)$ denotes stop-gradient. This prevents the contrastive objective from modifying the distributional representations solely to reduce the negative weights. We incorporate these negative weights into the negative terms of the

bidirectional contrastive loss based on the fused similarity $\mathbf{S}$. For notational simplicity, we set $\omega_{mm}^{t2v} = \omega_{mm}^{v2t} = 1$.

$$\mathcal{L}_{\text{UNC}} = -\frac{1}{2B}\sum_{m=1}^{B}\left[\log\frac{\exp(\mathbf{S}(m,m)/\tau)}{\sum_{n=1}^{B}\omega_{mn}^{t2v}\exp(\mathbf{S}(m,n)/\tau)} + \log\frac{\exp(\mathbf{S}(m,m)/\tau)}{\sum_{n=1}^{B}\omega_{nm}^{v2t}\exp(\mathbf{S}(n,m)/\tau)}\right]. \tag{15}$$

Finally, the overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{UNC}} + \lambda_{\text{PDM}}\mathcal{L}_{\text{PDM}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{16}$$

where λ_{PDM} and λ_{KL} control the weights of the two terms.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three commonly used datasets: How2Sign (Duarte et al., 2021), PHOENIX-2014T (Camgoz et al., 2018), and CSL-Daily (Zhou et al., 2021a). How2Sign contains 31,019 training, 1,738 validation, and 2,348 test video-text pairs. PHOENIX-2014T contains 7,096 training, 519 validation, and 642 test pairs, while CSL-Daily contains 18,401 training, 1,077 validation, and 1,176 test pairs. We use the official splits for training, validation, and testing.

Evaluation Metrics. We report text-to-video (T2V) and video-to-text (V2T) retrieval. Recall at K (R@1, R@5, and R@10; higher is better) measures the percentage of queries whose paired target is found among the top K candidates. Rsum is the sum of all recall values. We also adopt median rank (MedR) and mean rank (MnR). MedR denotes the median position of the ground-truth sample within the retrieved ranking list, while MnR corresponds to the average rank of the ground-truth.

Implementation Details. We evaluate MGUC under two visual-input settings: RGB-only and RGB+Pose. All ablation studies use RGB+Pose. The model is trained on 8 NVIDIA RTX 4090 GPUs. The dimension D is set to 512. Videos are sampled at 24 FPS with at most 64 clips, and text queries are truncated to 32 tokens. We train the model with AdamW for 200 epochs using a batch size of 128. The learning rate is set to 10^{-5}. We set $\lambda_{\text{PDM}} = 10^{-2}$, $\lambda_{\text{KL}} = 10^{-5}$, $\tau_a = 1.0$ and $\gamma = 0.8$. The model is trained with the traditional contrastive objective (Eq. 15, with $\omega = 1$) for the first 30% of training epochs to learn stable initial cross-modal representations.

4.2 Performance Evaluation

In Table 1, we provide a comprehensive comparison between our approach and existing sign-text retrieval baselines over three widely-used benchmarks, How2Sign, PHOENIX-2014T, and CSL-Daily. Our method achieves competitive performance compared with existing approaches on all three datasets for both with-pose and without-pose settings, which demonstrates the effectiveness of our method. Moreover, our method obtains higher Rsum scores across different datasets, further demonstrating the effectiveness of our model for sign-text retrieval.

As shown in Fig. 4, we visualize text-to-sign retrieval results of the proposed method. In all examples, our model retrieves the paired videos at Rank 1, demonstrating its ability to capture fine-grained action cues and sentence-level semantics. In the third example, it also retrieves a semantically related video at Rank 2, whose reference text conveys similar semantics. These examples suggest that our method captures both fine-grained correspondence and semantically related retrieval candidates. Additional examples of uncertain cases and corresponding retrieval results are provided in Fig. 11.

4.3 Ablation Study

The Effect of the Main Components. To validate the effectiveness of our core modules, we conduct ablation studies on How2Sign, as reported in Table 2. Removing FSM leads to a substantial performance drop, which reveals that the fine-grained interaction module is central to sign-text matching. When both FSM and PDM are discarded simultaneously, the retrieval results further degrade, demonstrating their complementary contributions.

Method	Visual Input		Text-to-Video (T2V)					Video-to-Text (V2T)					Rsum↑
	RGB	Pose	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	
How2Sign													
SA-SR (Duarte et al., 2022)	✓		18.9	32.1	36.5	22.0	–	11.6	27.4	32.5	24.0	–	159.0
SA-CM (Duarte et al., 2022)	✓		24.3	40.7	46.5	16.0	–	17.9	40.1	46.9	14.0	–	216.4
SA-COMB (Duarte et al., 2022)	✓		34.2	48.0	52.6	8.0	–	23.6	47.0	53.0	7.5	–	258.4
CiCo (Cheng et al., 2023)	✓		56.6	69.9	74.7	1.0	–	51.6	64.8	70.1	1.0	–	387.7
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	395.1
SEDs (Jiang et al., 2024a)	✓	✓	61.7	75.1	79.6	1.0	44.9	57.1	69.8	75.3	1.0	59.1	418.6
MGUC(ours)	✓		61.3	73.9	78.2	1.0	45.2	55.7	68.3	72.9	1.0	60.8	410.3
MGUC(ours)	✓	✓	**63.2**	**76.4**	**80.7**	1.0	**43.8**	**58.6**	**71.9**	**76.4**	1.0	**58.2**	**427.2**
PHOENIX-2014T													
SA-SR (Duarte et al., 2022)	✓		30.2	53.1	63.4	4.5	–	28.8	52.0	60.8	5.5	–	288.3
SA-CM (Duarte et al., 2022)	✓		48.6	76.5	84.6	2.0	–	50.3	78.4	84.4	1.0	–	422.8
SA-COMB (Duarte et al., 2022)	✓		55.8	79.6	92.1	1.0	–	53.1	79.4	86.1	1.0	–	446.1
CiCo (Cheng et al., 2023)	✓		69.5	86.6	92.1	1.0	–	70.2	88.0	92.8	1.0	–	499.2
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	509.9
SEDs (Jiang et al., 2024a)	✓	✓	75.4	91.6	**95.8**	1.0	3.5	76.3	92.7	**95.5**	**1.0**	3.1	527.3
MGUC(ours)	✓		75.1	90.9	94.4	1.0	4.0	74.9	92.0	95.0	**1.0**	4.2	522.3
MGUC(ours)	✓	✓	**77.3**	**92.1**	95.6	**1.0**	**3.1**	**77.4**	**93.3**	95.1	**1.0**	**3.0**	**530.8**
CSL-Daily													
CiCo (Cheng et al., 2023)	✓		75.3	88.2	91.9	1.0	–	74.7	89.4	92.2	1.0	–	511.7
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	518.4
SEDs (Jiang et al., 2024a)	✓	✓	84.5	94.1	95.5	1.0	3.3	84.1	94.0	95.8	1.0	3.4	548.0
MGUC(ours)	✓		82.1	91.3	94.0	1.0	4.4	81.9	92.6	94.6	1.0	4.7	536.5
MGUC(ours)	✓	✓	**86.5**	**95.8**	**97.1**	1.0	**2.7**	**89.1**	**95.5**	**97.1**	1.0	**2.5**	**561.1**

Table 1: Comparison with state-of-the-art methods on How2Sign, PHOENIX-2014T, and CSL-Daily.

Method	T2V R@1	V2T R@1	Rsum
Ours (Full)	**63.2**	**58.6**	**427.2**
w/o FSM	49.8	45.2	365.9
w/o PDM	60.4	55.2	408.3
w/o UACC	62.1	57.5	422.7
w/o FSM & PDM	46.7	41.9	334.1

Table 2: Ablation of the proposed components.

Method	T2V R@1	V2T R@1	Rsum
Mean–Max	58.8	55.3	414.5
Softmax weighting	60.2	56.1	418.1
Bi-SA	61.7	57.3	420.6
Bi-SA w/ video [Null]	62.5	57.9	424.4
Bi-SA w/ text [Null]	62.3	**58.6**	425.7
Bi-SA w/ dual [Null]	**63.2**	**58.6**	**427.2**

Table 3: Ablation of fine-grained matching designs.

Method	T2V R@1	V2T R@1	Rsum
Uncertainty head	60.2	55.8	410.1
Ours (global only)	62.1	57.7	422.9
Ours (local only)	61.7	57.5	420.7
Ours (both)	**63.2**	**58.6**	**427.2**

Table 4: Ablation of the probabilistic modeling.

Method	T2V R@1	V2T R@1	Rsum
KL divergence	45.3	45.3	324.2
JS divergence	44.3	46.3	326.4
2-Wasserstein distance	55.3	51.8	379.6
Expected likelihood kernel	60.7	56.8	419.5
Match probability	61.2	56.3	420.7
Bhattacharyya (ours)	**63.2**	**58.6**	**427.2**

Table 5: Comparison of distributional metrics.

Similarity Aggregation in Fine-Grained Matching. We conduct an in-depth ablation analysis of our fine-grained matching module in Table 3. Our method outperforms conventional similarity aggregation approaches, which lack interactions across different features. Furthermore, we thoroughly validate the effectiveness of our `[Null]` token, which suppress spurious matches between natural language and sign videos.

Global and Local Cues in Uncertainty Modeling. Table 4 evaluates probabilistic modeling with different token interactions. Global-only restricts the distributional tokens' direct attention to the global token, whereas local-only restricts it to local feature tokens. Using both interactions achieves the best performance, showing that global and local semantics are complementary. The uncertainty-head baseline predicts probabilistic distributions without direct interactions between feature tokens and yields inferior retrieval performance.

Distributional Metrics in Uncertainty Modeling. Table 5 compares different distributional metrics on How2Sign. Each metric replaces $d(m, n)$. Since the metrics differ in scale and direction, we transform their outputs into normalized distances, with smaller values indicating closer distributions, before computing the relative distance gaps and negative weights. The Bhattacharyya distance achieves the best retrieval performance, indicating that it provides an effective measure of distributional discrepancy for calibrating unpaired samples in sign–text matching.

Hyper-parameter Sensitivity. We analyze the effects of γ and λ_{PDM} on the How2Sign dataset in Fig. 5. The performance remains stable within a moderate range of γ and peaks at $\gamma = 0.8$. For λ_{PDM}, performance improves as the weight increases up to 10^{-2}, whereas it drops when a larger weight is assigned to the distributional objective. Consequently, we adopt $\gamma = 0.8$ and $\lambda_{\mathrm{PDM}} = 10^{-2}$ for all our experiments.



Figure 4: Qualitative visualization of text-to-sign retrieval. Under each video clip is its corresponding ground-truth text description.

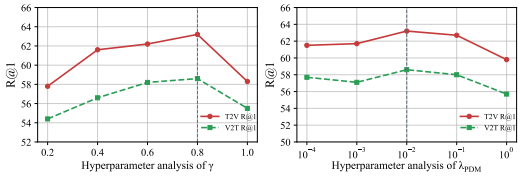

Figure 5: Sensitivity analysis of γ and λ_{PDM}.

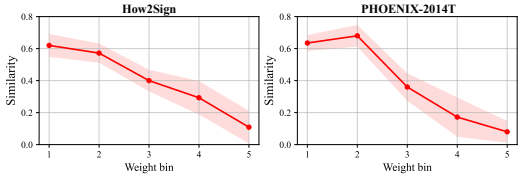

Figure 6: Semantic relatedness across weight bins.

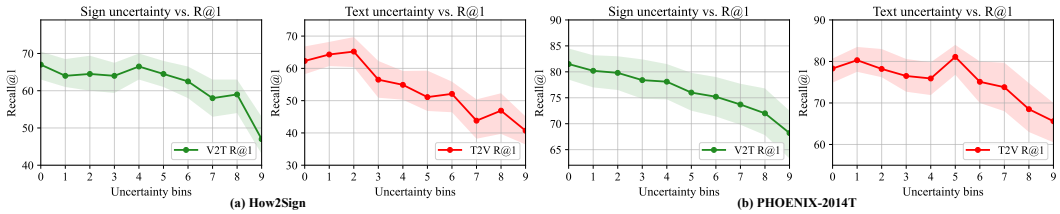

Figure 7: Retrieval performance across uncertainty bins. Sign and text queries are grouped into ten bins with increasing uncertainty on How2Sign and PHOENIX-2014T.

4.4 Analysis

Uncertainty vs. Retrieval Difficulty. We first examine whether the estimated uncertainty is associated with retrieval difficulty. We define the uncertainty of each query as $||\boldsymbol{\sigma}^2||_1$, and partition sign and text queries into ten bins via the uncertainty level. As shown in Fig. 7, R@1 generally decreases as uncertainty increases on both How2Sign and PHOENIX-2014T. The decrease with increasing uncertainty suggests that the learned uncertainty is associated with sign–text retrieval difficulty, supporting the need to account for matching uncertainty during training.

Negative Weights vs. Semantic Relatedness. We further examine whether the learned UACC weights are associated with the semantic relatedness of unpaired samples. For each unpaired pair (t_m, v_n), we group pairs according to the learned negative weight ω_{mn} and measure the semantic similarity between t_m and t_n using MPNet (Song et al., 2020). As shown in Fig. 6, lower-weight pairs exhibit higher textual semantic similarity on both datasets. This observation suggests that UACC tends to down-weight unpaired samples that are semantically related to the query, thereby reducing their influence as negative supervision. Furthermore, we provide a visual analysis in the appendix (Fig. 9 and 10) to investigate the contribution of individual words to the uncertainty.

5 Conclusion

In this paper, we propose MGUC, a multi-granular semantic matching framework with uncertainty-aware calibration for sign language retrieval. Coarse-grained global matching captures overall event semantics and preserves holistic sign–text consistency. Fine-grained local matching captures subtle motion details through interaction-aware bidirectional similarity aggregation, with learnable `[Null]` tokens mitigating spurious matches involving transitional clips and words without explicit sign gestures. To address uncertainty in sign–text matching, MGUC represents sign videos and texts as Gaussian distributions and uses Bhattacharyya distances to calibrate the negative contributions of unpaired pairs during contrastive learning. Experiments on How2Sign, PHOENIX-2014T, and CSL-Daily show consistent retrieval gains under different visual input settings.

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

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 (ECCV)*, pp. 35–53, 2020.

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10023–10033, 2020.

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5709–5718. IEEE, 2020.

Haipeng Chen, Yu Liu, Xun Yang, Yuheng Liang, and Yingda Lyu. Causality-aligned semantic recovery for incomplete cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20050–20058, 2026.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

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

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. pp. 1880–1891, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.

Jianfeng Dong, Xiaoman Peng, Zhe Ma, Daizong Liu, Xiaoye Qu, Xun Yang, Jixiang Zhu, and Baolong Liu. From region to patch: Attribute-aware foreground-background contrastive learning for fine-grained fashion retrieval. *arXiv preprint arXiv:2305.10260*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14074–14084. IEEE, 2022.

Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13723–13733, October 2023.

Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. Distilling cross-temporal contexts for continuous sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10771–10780, 2023.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11067–11076. IEEE, 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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision – ECCV 2018*. Springer International Publishing, 2018.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *ArXiv*, abs/2208.02515, 2022.

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. *ACM Multimedia*, 2022.

Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*, 2018.

Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4165–4174, 2019.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6902–6911, 2019.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *European Conference on Computer Vision*, pp. 53–70. Springer, 2020.

Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *International Conference on Computer Vision (ICCV)*, pp. 11946–11955, 2021.

Xiaohan Wang, Linchao Zhu, and Yi Yang. T2VLAD: global-local sequence alignment for text-video retrieval. *CoRR*, abs/2104.10054, 2021.

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.

Yang Yang, Haoyu Shi, and Huaiwen Zhang. Hierarchical semantics alignment for 3d human motion retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 1083–1092. ACM, 2024.

Shengwei Zhao, Linhai Xu, Yuying Liu, and S. Du. Multi-grained representation learning for cross-modal retrieval. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1316–1325. IEEE, 2021a.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. pp. 768–779, 2021b.

Minghao Zhu, Xiao Lin, Ronghao Dang, Chengju Liu, and Qijun Chen. Fine-grained spatiotemporal motion alignment for contrastive video representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 4725–4736. Association for Computing Machinery, 2023.

A Appendix

This supplementary material presents implementation details and additional experiments. We first describe the implementation details of MGUC, in Sec. A.1. We then report additional quantitative analyses in Sec. A.2, including the influence of inference-time post-processing and the computational cost of the proposed method. Next, we analyze the role of the Bhattacharyya-based distributional similarity in Sec. A.3. In Sec. A.4, we analyze the Bhattacharyya distances learned by probabilistic distribution modeling (PDM) and examine how uncertainty-aware contrastive calibration (UACC) uses the distance to adjust negative weights during training. We then analyze frequent words across text-uncertainty bins in Sec. A.5 and provide qualitative retrieval results for high-uncertainty queries in Sec. A.6. Finally, Sec. A.7 reports a human evaluation of the association between negative weights and human-assessed semantic relevance.

Meanwhile, we address the following questions and potential concerns:

- **RQ1:** Does the proposed model introduce substantial additional computational overhead or resource consumption?
- **RQ2:** Why is the distributional similarity not directly incorporated into the retrieval score?
- **RQ3:** Which words in sign-language queries are associated with higher predicted text uncertainty?

A.1 More Implementation Details

More Experimental Details. The offline RGB encoder is an I3D network pre-trained on BSL-1K (Albanie et al., 2020). For the Transformer encoder, we use 12 layers with a hidden size of 768 and initialize it from the CLIP ViT-B/32 image encoder (Radford et al., 2021). We modify the input 1D convolution to accommodate the 1024-dimensional I3D features, which are subsequently projected to 512 dimensions. For the RGB+Pose setting, we follow SEDS (Jiang et al., 2024a) to construct the pose stream and fuse the RGB and pose representations. Specifically, the online pose encoder consists of two GCN modules and a temporal convolution module, where the GCN modules for both hands are initialized with SignBERT (Hu et al., 2021). The pose features are projected to 512 dimensions, and the RGB-pose fusion strategy follows SEDS (Jiang et al., 2024a). For all datasets, sign videos are sampled at 24 FPS with at most 64 clips per video, and text queries are truncated to 32 tokens. We use the AdamW optimizer (Kingma & Ba, 2014) with a cosine warm-up learning-rate schedule and freeze the offline RGB encoder during training.

Implementation Details of Bi-SA. We use at most $K_v = 64$ visual clips and $K_t = 32$ text tokens. All clip-level and word-level features are projected to $D = 512$ dimensions. We append one learnable $[\mathrm{Null}]$ token of the same dimension to each modality. The resulting clip–word similarity matrix is $\mathbf{C} \in \mathbb{R}^{(K_v+1)\times(K_t+1)}$. The visual SA operator uses learnable parameters $\mathbf{W}_v \in \mathbb{R}^{(K_v+1)\times(K_v+1)}$ and $\mathbf{b}_v \in \mathbb{R}^{K_v+1}$, while the textual SA operator uses $\mathbf{W}_t \in \mathbb{R}^{(K_t+1)\times(K_t+1)}$ and $\mathbf{b}_t \in \mathbb{R}^{K_t+1}$. For an SA input vector $\mathbf{x}$ and its mask $\mathbf{m}$, we first compute the masked relevance distribution $\mathbf{p} = \mathrm{Softmax}(\mathbf{x}, \mathbf{m})$. We then compute $\mathbf{z} = \mathbf{W}\mathbf{p} + \mathbf{b}$ and obtain the aggregation weights $\boldsymbol{a} = \mathrm{Softmax}(\mathbf{z}, \mathbf{m})$. Padded positions are masked before both softmax operations. The $[\mathrm{Null}]$ position remains valid and is therefore included in both softmax operations and in the learnable interaction $\mathbf{W}\mathbf{p} + \mathbf{b}$. However, it is excluded from the final weighted sum, and the remaining weights are not renormalized after this exclusion. Consequently, assigning a large weight to $[\mathrm{Null}]$ reduces the contribution of valid clip–word similarities and allows unmatched clips or words to be suppressed. The complete computation is summarized in Algorithm 1. Bi-SA applies the two SA operators in both directions. In the visual-to-text direction, SA_v is applied independently to each column of $\mathbf{C}$, and the resulting vector is aggregated by SA_t. In the text-to-visual direction, SA_t is applied independently to each row of $\mathbf{C}$, and the resulting vector is aggregated by SA_v. The final local similarity is computed as $S_{\mathrm{loc}} = \frac{1}{2}(S_{v\to t} + S_{t\to v})$.

Algorithm 1 Bidirectional Interactive Similarity Aggregation

Require: Similarity matrix $\mathbf{C} \in \mathbb{R}^{(K_v+1)\times(K_t+1)}$; visual mask $\mathbf{m}_v$; text mask $\mathbf{m}_t$
Ensure: Fine-grained similarity S_{loc}

function $\mathrm{SA}(\mathbf{x}, \mathbf{m}, \mathbf{W}, \mathbf{b})$
  $\mathbf{p} \leftarrow \mathrm{Softmax}(\mathbf{x}, \mathbf{m})$ — compute the initial relevance distribution
  $\mathbf{z} \leftarrow \mathbf{W}\mathbf{p} + \mathbf{b}$ — model interactions among positions
  $\mathbf{a} \leftarrow \mathrm{Softmax}(\mathbf{z}, \mathbf{m})$ — compute the final aggregation weights
  $s \leftarrow \sum_{k \in \mathcal{V}(\mathbf{m})} a_k x_k$ — aggregate valid positions and exclude $[\mathrm{Null}]$
  return s
end function
for $i = 1, \ldots, K_t + 1$ **do**
  $r_i^{v\to t} \leftarrow \mathrm{SA}(\mathbf{C}_{:,i}, \mathbf{m}_v, \mathbf{W}_v, \mathbf{b}_v)$ — aggregate clip similarities for each word
end for
$\mathbf{r}^{v\to t} \leftarrow (r_1^{v\to t}, \ldots, r_{K_t+1}^{v\to t})$
$s_{v\to t} \leftarrow \mathrm{SA}(\mathbf{r}^{v\to t}, \mathbf{m}_t, \mathbf{W}_t, \mathbf{b}_t)$ — compute the visual-to-text similarity
for $j = 1, \ldots, K_v + 1$ **do**
  $r_j^{t\to v} \leftarrow \mathrm{SA}(\mathbf{C}_{j,:}, \mathbf{m}_t, \mathbf{W}_t, \mathbf{b}_t)$ — aggregate word similarities for each clip
end for
$\mathbf{r}^{t\to v} \leftarrow (r_1^{t\to v}, \ldots, r_{K_v+1}^{t\to v})$
$s_{t\to v} \leftarrow \mathrm{SA}(\mathbf{r}^{t\to v}, \mathbf{m}_v, \mathbf{W}_v, \mathbf{b}_v)$ — compute the text-to-visual similarity
$S_{\mathrm{loc}} \leftarrow \frac{1}{2}(s_{v\to t} + s_{t\to v})$ — fuse the two directional similarities
return S_{loc}

Method	Params	FLOPs	Infer. time	Rsum
SEDS	213.97M	4.98G	6.54ms	418.6
MGUC (ours)	217.39M	5.50G	7.72ms	427.2

Table 6: Computational cost comparison between SEDS and MGUC on How2Sign. FLOPs and inference time are measured per query.

Method	Visual Input		Text-to-Video (T2V)			Video-to-Text (V2T)			Rsum↑
	RGB	Pose	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
MGUC (Ours)	✓		61.3	73.9	78.2	55.7	68.3	72.9	410.3
+ DSL	✓		**62.0** (+0.7)	**74.5** (+0.6)	**79.0** (+0.8)	**56.8** (+1.1)	**69.4** (+1.1)	**73.7** (+0.8)	**415.4** (+5.1)
MGUC (Ours)	✓	✓	63.2	76.4	80.7	58.6	71.9	76.4	427.2
+ DSL	✓	✓	**64.3** (+1.1)	**77.0** (+0.6)	**81.2** (+0.5)	**59.8** (+1.2)	**72.8** (+0.9)	**77.1** (+0.7)	**432.2** (+5.0)

Table 7: Effect of DSL during inference on How2Sign.

A.2 More Quantitative Results

Computational Cost Comparison. (RQ1) We compare our proposed MGUC with SEDS, the state-of-the-art baseline. Both models are evaluated under the same hardware, input sampling strategy, batch size, and feature dimensions. FLOPs and inference time are measured per query. As shown in Table **??**, MGUC uses 217.39M parameters and requires 5.50G FLOPs per query, compared with 213.97M parameters and 4.98G FLOPs for SEDS. The corresponding inference time increases only slightly, from 6.54 ms to 7.72 ms per query, indicating a modest additional computational cost. This overhead is mainly introduced by the bidirectional similarity aggregation module. In exchange, MGUC improves the How2Sign Rsum from 418.6 to 427.2, yielding an absolute improvement of 8.6 points. The default inference score does not use UACC negative weights; UACC adjusts the negative contributions only during training.

Effect of Post-Processing. We further investigate the effect of Dual Softmax (DSL) (Cheng et al., 2021) as an inference-time post-processing strategy. Specifically, DSL is applied to the fused similarity matrix to refine the bidirectional similarity scores without additional training. As illustrated in Tables **??**, **??**, and **??**, DSL improves retrieval performance across all evaluated datasets and visual input settings. The improvements are observed for both text-to-video and video-to-text retrieval, indicating that DSL can provide complementary score refinement for the proposed multi-granular matching framework.

A.3 Discussion on Distributional Similarity **(RQ2)**

We investigate whether the Bhattacharyya-based distributional similarity should be directly incorporated into the inference score. As shown in Table **??**, directly adding distributional similarity consistently degrades retrieval performance on all three datasets. The fused retrieval score is optimized to rank paired targets by jointly modeling global semantics and fine-grained clip–word correspondences. By contrast, Bhattacharyya-based distributional similarity may assign high values to multiple semantically related candidates, including unpaired ones. Directly adding it to the inference score can therefore reduce discrimination under the single-pair retrieval protocol. We use this similarity as an auxiliary training signal for negative-weight calibration rather than as a ranking score. The distribution-level objective learns pairwise distributional relations, and the resulting distance gaps are used by UACC to adjust the negative contributions of unpaired samples. Thus, the probabilistic branch supports uncertainty-aware negative calibration without replacing the deterministic retrieval score.

A.4 Analysis of Bhattacharyya Distances and Negative-Weight Calibration

We examine whether the learned Gaussian representations distinguish paired samples, semantically related unpaired samples, and clearly mismatched unpaired samples. Let (t_m, v_m) denote a paired text–video sample. For the same text query t_m, we consider a semantically related but unpaired

Method	Visual Input		Text-to-Video (T2V)			Video-to-Text (V2T)			Rsum↑
	RGB	Pose	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
MGUC (Ours)	✓		75.1	90.9	94.4	74.9	92.0	95.0	522.3
+ DSL	✓		**75.9** (+0.8)	**91.3** (+0.4)	**94.7** (+0.3)	**76.0** (+1.1)	**92.6** (+0.6)	**95.5** (+0.5)	**526.0** (+3.7)
MGUC (Ours)	✓	✓	77.3	92.1	95.6	77.4	93.3	95.1	530.8
+ DSL	✓	✓	**77.9** (+0.6)	**92.4** (+0.3)	**95.8** (+0.2)	**78.3** (+0.9)	**93.8** (+0.5)	**95.5** (+0.4)	**533.7** (+2.9)

Table 8: Effect of DSL during inference on PHOENIX-2014T.

Method	Visual Input		Text-to-Video (T2V)			Video-to-Text (V2T)			Rsum↑
	RGB	Pose	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
MGUC (Ours)	✓		82.1	91.3	94.0	81.9	92.6	94.6	536.5
+ DSL	✓		**83.6** (+1.5)	**92.2** (+0.9)	**94.5** (+0.5)	**83.5** (+1.6)	**93.4** (+0.8)	**95.0** (+0.4)	**542.2** (+5.7)
MGUC (Ours)	✓	✓	86.5	95.8	97.1	89.1	95.5	97.1	561.1
+ DSL	✓	✓	**87.7** (+1.2)	**96.4** (+0.6)	**97.4** (+0.3)	**90.8** (+1.7)	**96.3** (+0.8)	**97.5** (+0.4)	**566.1** (+5.0)

Table 9: Effect of DSL during inference on CSL-Daily.

Inference score	**How2Sign**			**PHOENIX-2014T**			**CSL-Daily**		
	T2V R@1	V2T R@1	Rsum	T2V R@1	V2T R@1	Rsum	T2V R@1	V2T R@1	Rsum
Ours (default)	63.2	58.6	427.2	77.3	77.4	530.8	86.5	89.1	561.1
+ Distributional similarity	56.9	53.2	408.7	66.2	65.4	475.0	76.4	78.3	503.4

Table 10: Effect of directly incorporating the Bhattacharyya-based distributional similarity during inference. S denotes the original fused retrieval similarity. Directly adding distributional similarity reduces retrieval performance, supporting its use as an auxiliary training signal for uncertainty-aware negative calibration rather than as a ranking score.

video v_{rel} and a clearly mismatched video v_{mis}. We expect their Bhattacharyya distances to follow

$$d(t_m, v_m) < d(t_m, v_{\text{rel}}) < d(t_m, v_{\text{mis}}),$$

where $d(\cdot, \cdot)$ denotes the Bhattacharyya distance defined in the main text.

To select the two types of unpaired samples, we use a frozen MPNet text encoder (Song et al., 2020) to rank candidate text descriptions by their cosine similarity to the query. We select the highest-ranked non-paired descriptions as semantically related samples and the lowest-ranked ones as clearly mismatched samples. We then examine their Bhattacharyya distances, distance gaps relative to the paired sample, and assigned negative weights.

For ease of comparison, we also report the Bhattacharyya coefficient $\exp(-d)$ derived from each distance d. A larger coefficient indicates a smaller distance and thus closer distributions. Table **??** reports these coefficients, the distance gaps Δ, and the negative weights ω. The paired sample has the smallest distance and the largest coefficient. Among unpaired samples, the semantically related sample has a smaller distance gap and a larger coefficient than the clearly mismatched sample. Consistent with these distance patterns, the proposed calibration assigns a lower negative weight to the semantically related sample (0.65) than to the clearly mismatched sample (0.92). This indicates that relative Bhattacharyya distances can guide the adjustment of unpaired samples' negative contributions. Fig. 8 tracks the Bhattacharyya coefficients and negative weights during training. Paired samples have the smallest distances, followed by semantically related and clearly mismatched samples. The negative weights likewise distinguish the two types of unpaired samples, assigning lower weights to semantically related samples. These results show that the learned distributional representations support adaptive negative weighting during contrastive learning.

A.5 Frequent Words across Text Uncertainty Bins (RQ3)

We further analyze the words associated with the learned text uncertainty. For each dataset, we partition text queries into four uncertainty bins in ascending order and identify the most relevant words within each bin. We use term frequency–inverse document frequency (TF-IDF) to identify

Pair type	$\exp(-d)$	Δ	ω
Paired sample	0.94	–	1.00
Semantic-related negative	0.71	0.28	0.65
Clearly mismatched negative	0.10	2.24	0.92

Table 11: Bhattacharyya coefficients $\exp(-d)$, relative distance gaps, and negative weights on How2Sign.

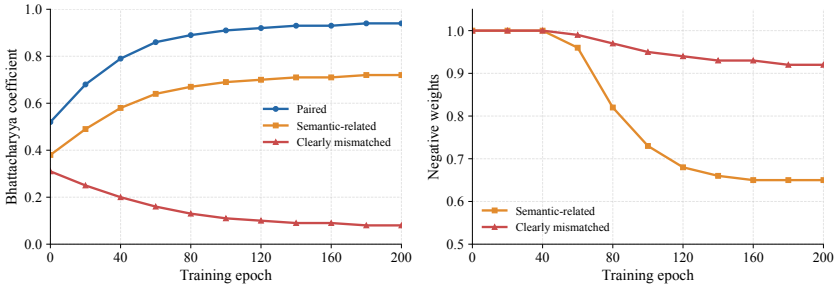

Figure 8: Training dynamics of Bhattacharyya coefficients and negative weights. We track paired, semantically related, and clearly mismatched samples during training.

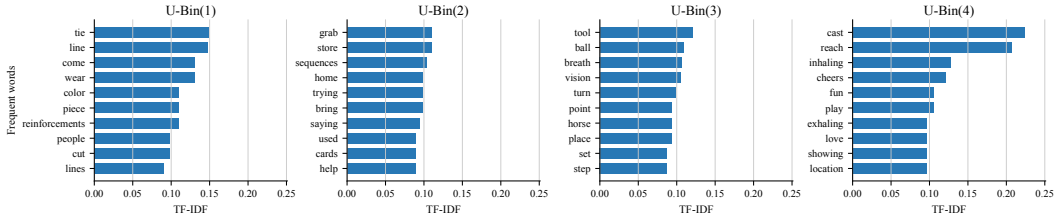

Figure 9: Frequent words in each uncertainty bin on How2Sign. Text queries are partitioned into four bins in ascending order of learned uncertainty. Each subplot shows the top-ranked words according to their TF-IDF scores within the corresponding bin.

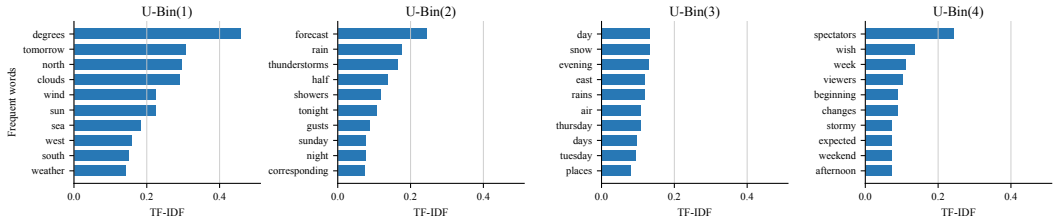

Figure 10: Frequent words in each uncertainty bin on PHOENIX-2014T. Text queries are partitioned into four bins in ascending order of learned uncertainty. Each subplot shows the top-ranked words according to their TF-IDF scores within the corresponding bin.

frequent words in each uncertainty bin, defined as

$$\text{TF-IDF}(i) = (1 + \log n_i) \log \frac{N}{n_i}, \tag{17}$$

where N is the total number of text queries and n_i is the number of queries containing word i. We use the ground-truth text descriptions to compute the TF-IDF scores.

Figures 9 and 10 show the top-ranked words in each uncertainty bin on How2Sign and PHOENIX-2014T, respectively. This provides an interpretable view of the linguistic content captured by the learned uncertainty representations.



Figure 11: Qualitative text-to-video retrieval results for queries with relatively high learned uncertainty. For each query, we show its text uncertainty $\|\boldsymbol{\sigma}_t^2\|_1$ and the top-3 retrieved sign videos with their corresponding text descriptions.

A.6 Visualization of High-Uncertainty Retrieval Cases

We further provide qualitative examples to examine the learned uncertainty in text-to-video retrieval. Fig. 11 presents the top-3 retrieved sign videos for text queries with relatively high predicted uncertainty, measured by $\|\sigma_t^2\|_1$. Several non-paired retrieved videos remain semantically related to their queries, showing that single-pair annotations may not capture all plausible retrieval matches. Such samples motivate reducing their negative contributions during contrastive learning.

A.7 Human Evaluation

To directly examine whether UACC assigns lower negative weights to semantically related unpaired pairs, we conduct a human evaluation on the official How2Sign test split. We randomly sampled 20 text queries. For each query, we construct a candidate pool from the top-10 videos retrieved by MGUC, MGUC without UACC, and SEDS. We additionally include the ground-truth paired video and one randomly sampled non-paired video from the same test split. After removing duplicate videos within each candidate pool, the evaluation set contains 410 video–text pairs. The annotators are not shown the retrieval method, ranking position, similarity score, or UACC negative weight.

Each video–text pair is independently evaluated by two annotators proficient in Sign Language. The annotators assigned one of four relevance levels:

- **0**: unrelated;
- **1**: locally or visually similar, but different in overall meaning;
- **2**: semantically related and a plausible retrieval result;
- **3**: fully relevant or an exact semantic match.

Analysis	Spearman ρ	AUROC	AUPRC
$1-\omega$ vs. human relevance	0.46	0.78	0.71

Table 12: Association between UACC negative weights and human-assessed relevance on unpaired video–text pairs. A larger value of $1-\omega$ indicates greater down-weighting of a sample as a negative.

Disagreements were resolved through discussion between the two annotators. Inter-annotator agreement, measured using Cohen's κ, was 0.83. We examine whether UACC's negative weights are associated with human-assessed semantic relevance. This analysis are restricted to unpaired pairs because the weight of each paired sample are fixed to one. We use $1-\omega_{mn}$ as a down-weighting score, where a larger value indicates that the sample contributes less to negative supervision. We calculate the Spearman correlation between this score and the ordinal human relevance score. We also evaluate whether it distinguishes semantically related pairs (relevance levels 2 or 3) from unrelated or only locally similar pairs (levels 0 or 1). As shown in Table 12, the down-weighting score is moderately correlated with human-assessed relevance ($\rho = 0.46$), achieving an AUROC of 0.78 and an AUPRC of 0.71. These results indicate that samples receiving lower negative weights are more likely to be judged semantically relevant by human annotators.

A.8 Limitations and Future Work

Despite the consistent improvements achieved across three sign language retrieval benchmarks, our evaluation is still constrained by the annotation protocol of existing datasets. These benchmarks typically provide only one annotated video–text pair for each query and do not exhaustively label semantically related alternatives. As a result, standard paired-retrieval metrics may treat plausible but unpaired matches as negatives. The MPNet-based analysis and human evaluation provide supporting evidence that some unpaired samples are semantically relevant, but they do not constitute exhaustive relevance annotations. In particular, the human evaluation is limited to a subset of the How2Sign test set. Future work will explore benchmarks with multi-positive annotations and graded relevance labels, together with broader human evaluation protocols. Such annotations would enable more direct assessment of uncertainty-aware negative weighting and the use of multi-positive retrieval metrics, such as multi-positive Recall, mAP, and R-Precision.