Semantic Hardness Is Not Visual Hardness: Sign-Aware Hard Negative Mining for Sign Language Retrieval

Junmyeong Lee[1] **Chan Hur**[4] **ChangSu Choi**[2] **Sukmin Cho**[1]
Fitsum Gaim[1] **Eui Jun Hwang**[1] **Hoyun Song**[3*] **KyungTae Lim**[2,3*]
School of Computing[1] Graduate School of Culture Technology[2]
InnoCORE PRISM-AI Center[3] Korea Advanced Institute of Science and Technology[1,2,3]
ETRI Medical Informatics Laboratory[4]
david516@kaist.ac.kr chanhur@etri.re.kr
{choics2623,nelllpic,fitsum.gaim,ehwa20,hysong,ktlim}@kaist.ac.kr

Abstract

Sign Language Retrieval (SLRet) enables efficient access to sign language content but remains fragile in fine-grained scenarios where visually similar signs must be distinguished. We show that this limitation does not stem from model capacity, but from ineffective hard negative supervision. Specifically, we formulate fine-grained retrieval failures as a negative distribution mismatch: semantically distinct yet visually confusable signs are rarely treated as hard negatives, while existing text-based mining strategies fail to capture such visual ambiguity. To address this issue, we propose Sign-Aware Hard Negative Mining (SAN), which constructs hard negatives based on visual confusability in the sign embedding space rather than linguistic similarity. Experiments on PHOENIX-2014T demonstrate that SAN substantially improves fine-grained retrieval performance while preserving coarse-grained accuracy, highlighting the importance of aligning negative supervision with visual ambiguity in sign language retrieval. Code is available at Github repository.[1].

1 Introduction

Sign languages are the primary means of communication for the Deaf community, expressed through hand, body, and facial movements. The unique grammar and visual complexity of sign languages often create communication barriers between signers and non-signers. To bridge this gap, prior research has focused on sign language understanding, particularly sign language recognition (Hu et al., 2021; Jiang et al., 2021; Zuo et al., 2023) and translation (Camgöz et al., 2020; Zhou et al., 2023; Gong et al., 2024). However, the scarcity of sign language data leads to high error rates across both tasks (Cheng et al., 2023).

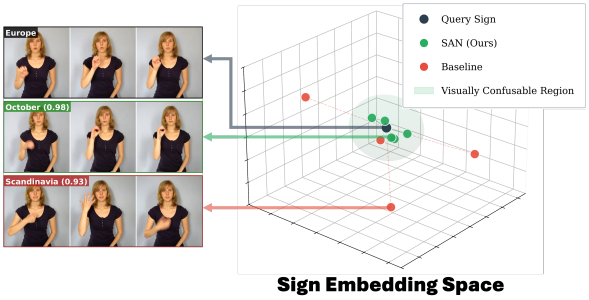


Figure 1: Illustration of fine-grained ambiguity in sign language retrieval. Semantically distinct words often correspond to visually similar signs, forming true hard negatives. SAN effectively targets these visually confusable instances that text-based mining methods frequently fail to address.

Recently, Sign Language Retrieval (SLRet) (Duarte et al., 2022; Cheng et al., 2023; Wu et al., 2024) has emerged as a promising task. SLRet aims to retrieve relevant sign language videos or texts from a database given a query. It enables efficient access to sign language content and facilitates the use of unannotated sign videos, helping mitigate the data scarcity of sign language resources (Duarte et al., 2022).

However, we identify a core bottleneck that current SLRet models fail to address. Since sign languages construct meaning within a restricted visual space, even subtle differences in hand shape, position, or trajectory can alter meaning—a phenomenon known as *sign confusability* (Albanie et al., 2020; Zuo et al., 2023). While existing models remain stable in coarse-grained retrieval, their performance degrades sharply in fine-grained retrieval scenarios where distinguishing these subtle distinctions is required. We argue that this degradation stems not from a lack of model capacity, but rather from training supervision that does not sufficiently expose the model to such cases.

We formulate the cause of this failure as a *negative distribution mismatch* in contrastive learning, which is exacerbated in sign language retrieval due

*Corresponding Author
[1] https://github.com/joonmy/SAN.git

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28262–28277

to frequent sign confusability. Importantly, this mismatch manifests in two forms. First, sign language datasets remain primarily coarse-grained (Joze and Koller, 2019; Rastgoo et al., 2021), lacking sufficient naturally occurring hard negatives to support fine-grained feature learning. Second, while previous studies have attempted to alleviate this limitation by generating hard negative captions through text-based perturbations (Momeni et al., 2023; Chen et al., 2024; Zhang et al., 2024), such approaches may still provide insufficient guidance for resolving sign confusability, as they rely on linguistic semantics rather than visual similarity.

Figure 1 intuitively illustrates this discrepancy. For a query word *Europe*, traditional text-based approaches select semantically related words like *Scandinavia* as hard negatives. However, because their corresponding signs are visually distinct, they serve merely as easy negatives for the model. By contrast, words like *October*, which are semantically unrelated but visually similar in their sign form, act as true hard negatives that the model struggles to distinguish. Yet such visually confusable samples are not covered by text-based negative mining approaches. Consequently, existing approaches fail to expose the model to learn the fine-grained differences located in the *Visually Confusable Region*.

Motivated by this observation, we propose **Sign-Aware Hard Negative Mining (SAN)**, which redefines the criteria for hard negative mining from linguistic similarity to visual confusability. SAN operates by (1) extracting high-confidence sign–word correspondences, (2) identifying signs that cluster closely in the sign language embedding space yet are semantically distinct, and (3) constructing hard negative captions using the words corresponding to these visually confusable signs, thereby correcting biases in the supervision signals.

We evaluate SAN on the PHOENIX-2014T dataset (Camgöz et al., 2018) and present three key findings. First, existing SLRet models exhibit substantial performance degradation in fine-grained retrieval scenarios involving visually confusable queries. Second, text-based negative mining fails to adequately address this problem. Third, SAN significantly improves fine-grained retrieval performance while preserving coarse-grained accuracy. These findings suggest that fine-grained discrimination failures in sign language retrieval stem from a supervision mismatch, and that aligning negatives with visual ambiguity is key to resolving it.

Our contributions are summarized as follows:

- We formulate fine-grained retrieval failures in sign language retrieval as a *negative distribution mismatch* problem.
- We propose Sign-Aware Hard Negative Mining (SAN), which aligns hard negatives with visual confusability in sign language.
- We demonstrate that SAN substantially improves fine-grained retrieval while maintaining stable coarse-grained performance.

2 Related Work

2.1 Sign Language Retrieval

Early research on sign language understanding primarily focused on Sign Language Recognition (SLR) (Zuo et al., 2023; Zhao et al., 2023; Ahn et al., 2024; Guan et al., 2025) and Sign Language Translation (SLT) (Zhang et al., 2023; Gong et al., 2024; Zhou et al., 2023; Hwang et al., 2025). SLR aims to predict glosses from sign language videos, whereas SLT focuses on generating natural language sentences.

More recently, Sign Language Retrieval (SLRet) has emerged as an important direction in sign language research. SLRet aims to retrieve semantically relevant videos or captions from sign language datasets given a query. This capability is particularly important given the rapid growth of online sign language content as it enables efficient access to sign resources (Duarte et al., 2022; Cheng et al., 2023). SPOT-ALIGN (Duarte et al., 2022) pioneered the SLRet task by introducing global alignment between sign videos and textual captions. Subsequently, CiCo (Cheng et al., 2023) advanced this paradigm by redefining SLRet as a cross-lingual retrieval problem and introducing cross-lingual contrastive learning (CLCL) to achieve finer alignment between sign units and textual tokens. SEDS (Jiang et al., 2024) further enhanced this approach by incorporating an additional keypoint modality for joint intra- and inter-modal learning. UPRet (Wu et al., 2024) addressed sign ambiguity by modeling both modalities as Gaussian distributions and aligning them via optimal transport.

However, a critical limitation persists: these methods predominantly optimize for coarse-grained retrieval, where in-batch negatives are relatively easy to distinguish. As a result, current models remain limited in fine-grained retrieval settings,

where subtle visual distinctions between signs are crucial. Despite its importance, this aspect has been largely underexplored, and we therefore focus on this challenge.

2.2 Hard Negatives in Vision–Language Models

Hard negatives provide informative supervision for contrastive learning by exposing models to hard-to-distinguish examples. In vision–language research, a common strategy to achieve this is to construct *hard negative captions*. Most prior works generate such negatives by modifying specific words in the original captions within the *textual domain*. To this end, standard approaches utilize rule-based heuristics (e.g., word swapping) (Yüksekgönül et al., 2023; Paiss et al., 2023; Chen et al., 2024) or leverage pretrained language models to synthesize counterfactual captions (Doveh et al., 2023; Momeni et al., 2023; Zhang et al., 2024; Patel et al., 2024). These methods implicitly assume that semantically difficult negatives (e.g., changing 'man' to 'boy') act as effective training signals for fine-grained visual discrimination.

While text-driven negative mining has proven effective in general vision–language domains, it is often suboptimal for sign language. In sign language, semantically related words are often expressed through entirely different motions, whereas semantically unrelated words can share high visual similarity (Zuo et al., 2023). Furthermore, because sign language conveys meaning through a constrained set of gestures, visually similar signs occur frequently (Albanie et al., 2020). As a result, text-based mining methods—operating solely in the textual semantic space—tend to produce *visually easy* negatives, failing to provide informative supervision for fine-grained visual discrimination.

To bridge this gap, we propose a sign-aware hard negative mining method that explicitly accounts for visual similarity of signs, identifying truly confusable negatives beyond textual semantics.

3 Methodology

In this section, we first introduce the preliminaries of sign language retrieval task in Section 3.1. We then present our proposed Sign-Aware Hard Negative Mining framework in Section 3.2.

3.1 Preliminary

Sign language retrieval aims to learn a shared embedding space that aligns sign videos $\mathcal{V}$ and textual descriptions $\mathcal{T}$ for cross-modal matching. We consider two standard settings: text-to-video (T2V) and video-to-text (V2T) retrieval, which retrieve the most relevant sign video $v \in \mathcal{V}$ for a given text query and the corresponding text $t \in \mathcal{T}$ for a sign video query, respectively.

Following prior work, we adopt the cross-lingual contrastive learning (CLCL) objective (Cheng et al., 2023) to capture fine-grained alignments between sign units and word tokens. Given a mini-batch of B video–text pairs $\{(v_i, t_i)\}_{i=1}^{B}$, we encode sign videos and texts using a visual encoder F and a text encoder G, producing sign-level and word-level feature sequences $v_i' = [s_i^0, s_i^1, \ldots, s_i^N]$ and $t_i' = [w_i^0, w_i^1, \ldots, w_i^M]$, where N and M denote the number of sign clips and words in the text, respectively. We then compute the sign–word similarity matrix and apply attention-weighted aggregation to obtain video–text similarity scores S_{V2T} and S_{T2V} for all video–text pairs in the mini-batch.

We optimize the retrieval model using the InfoNCE loss (Gutmann and Hyvärinen, 2010):

$$\mathcal{L}_{V2T} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{V2T}(v_i, t_i)/\tau)}{\sum_{j=1}^{B} \exp(S_{V2T}(v_i, t_j)/\tau)}, \quad (1)$$

$$\mathcal{L}_{T2V} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{T2V}(v_i, t_i)/\tau)}{\sum_{j=1}^{B} \exp(S_{T2V}(v_j, t_i)/\tau)}, \quad (2)$$

$$\mathcal{L}_{coarse} = \mathcal{L}_{V2T} + \mathcal{L}_{T2V}, \quad (3)$$

where τ is a learnable temperature parameter.

3.2 Sign-Aware Hard Negative Mining

To address the mismatch between the negative supervision signals and the visual ambiguity inherent in sign language, we propose *Sign-Aware Hard Negative Mining (SAN)*. From a contrastive learning perspective, SAN re-aligns the negative sample distribution with visually confusable regions of the sign space, which are underrepresented by standard mini-batch negatives. SAN consists of three components: (1) reliable sign–word pair mining, (2) visually similar negative word mining, and (3) hard negative caption generation.

Reliable Sign–Word Pair Mining To identify visually grounded hard negatives, we first extract reliable sign–word correspondences using a pretrained sign language retrieval model. For each video–text pair (v_i, t_i), we compute the sign–word similarity matrix

$$E^{(i)} = v_i' \cdot (t_i')^\top \in \mathbb{R}^{N \times M}. \quad (4)$$

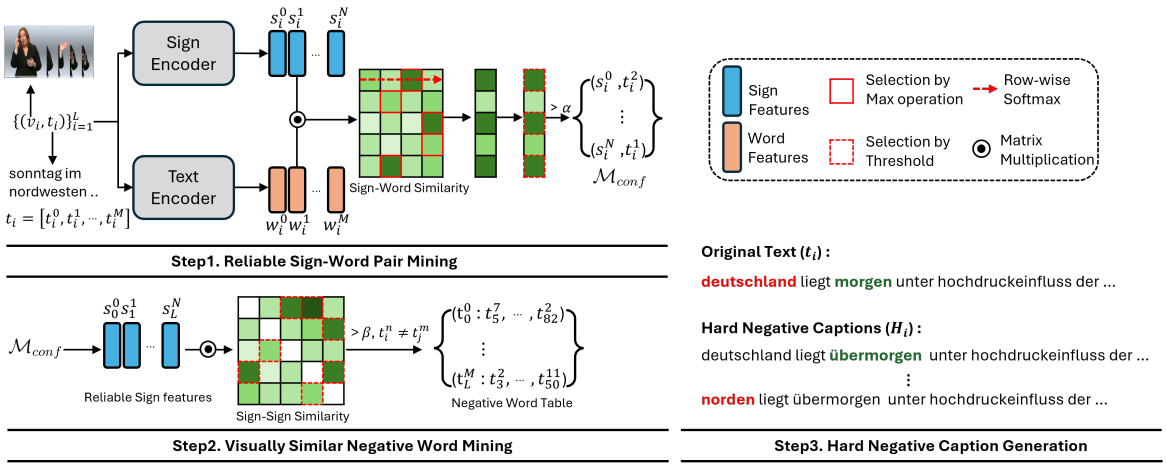


Figure 2: Overview of the Sign-Aware Hard Negative Mining (SAN) framework. SAN consists of three steps: (1) Reliable Sign–Word Pair Mining to extract high-confidence alignments; (2) Visually Similar Negative Word Mining to identify visually confusable but semantically distinct words; and (3) Hard Negative Caption Generation to construct challenging training samples via keyword substitution.

Subsequently, we normalize $E^{(i)}$ into a probability matrix via row-wise softmax:

$$P^{(i)}_{n,m} = \frac{\exp(E^{(i)}_{n,m}/\tau)}{\sum_{j=1}^{M} \exp(E^{(i)}_{n,j}/\tau)}. \quad (5)$$

For each sign token s^n_i, we take the maximum alignment probability:

$$p^n_i = \max_{1 \le m \le M} P^{(i)}_{n,m}. \quad (6)$$

Only matches with probabilities exceeding a threshold α are retained, yielding the set of reliable sign–word pairs:

$$\mathcal{M}_{\text{conf}} = \{(s^n_i, t^k_i) \mid p^n_i > \alpha,\ k = \arg\max_m P^{(i)}_{n,m}\}. \quad (7)$$

Note that this step is not intended to produce perfect alignments, but rather to conservatively filter out noisy alignments and stabilize the subsequent negative mining process.

Visually Similar Negative Word Mining Given $\mathcal{M}_{\text{conf}}$, we identify visually similar but semantically distinct signs to construct hard negative candidates. For each $(s^n_i, t^k_i) \in \mathcal{M}_{\text{conf}}$, we search for other pairs (s^m_j, t^l_j) whose sign features are visually similar by measuring:

$$\text{sim}(s^n_i, s^m_j) = \frac{s^n_i \cdot s^m_j}{\|s^n_i\| \|s^m_j\|}, \quad j \neq i. \quad (8)$$

We retain candidates whose similarity exceeds a threshold β and whose word tokens differ:

$$\mathcal{N}(t^k_i) = \{t^l_j \mid \text{sim}(s^n_i, s^m_j) > \beta,\ t^l_j \neq t^k_i\}. \quad (9)$$

By selecting negatives based on visual proximity between signs, this step reflects the constrained and structured articulation space of sign language, where visual similarity, rather than linguistic relatedness, governs confusability.

Hard Negative Caption Generation Using the mined candidate sets $\mathcal{N}$, we generate hard negative captions via keyword substitution. For each original caption t_i, we identify words with non-empty $\mathcal{N}(t)$, randomly select such words, and replace them with samples drawn from their corresponding candidate sets, yielding a set of hard negatives H_i.

We incorporate these negatives into contrastive learning via the following objective:

$$\mathcal{L}_{fine} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S_{V2T}(v_i, t_i)/\tau)}{\sum_{x \in H_i \cup \{t_i\}} \exp(S_{V2T}(v_i, x)/\tau)}. \quad (10)$$

Here, the number of substituted words and hard negatives serve as design choices, and their impact is systematically analyzed in Section 5.

Final Objective The final training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{coarse} + \lambda \cdot \mathcal{L}_{fine}, \quad (11)$$

where λ balances coarse-grained alignment and sign-aware hard negative supervision. This formulation allows us to isolate the effect of SAN while preserving the original retrieval objective, enabling controlled analysis in subsequent experiments.

4 Experiments

4.1 Experimental Setup

Dataset and Evaluation Goal We conduct experiments on **PHOENIX-2014T** (Camgöz et al., 2018), a standard benchmark for German Sign Language consisting of 7,096 training, 519 validation, and 642 test video–text pairs collected from TV weather forecast broadcasts. Beyond standard retrieval accuracy, our goal is to evaluate a model's ability to resolve *visually confusable signs*, which frequently arise in sign language due to its constrained articulatory space, yet are not explicitly captured by existing benchmarks.

Coarse-grained and Fine-grained Evaluation We consider two complementary evaluation settings. **Coarse-grained retrieval** utilizes the original test split and measures overall video–text matching performance under standard conditions, but does not probe whether a model can distinguish signs that differ only in subtle motion details. To explicitly assess this capability, we introduce a **fine-grained evaluation setting**, designed as a diagnostic *stress test* on visually ambiguous regions of the dataset.

Fine-grained Stress Test Construction The fine-grained test set is constructed via *single-word substitution*, producing minimally perturbed but semantically incorrect captions. Target words are restricted to those exhibiting visual confusability—specifically, words whose corresponding signs have at least one visually similar counterpart (cosine similarity $> \beta$ in the sign embedding space) among semantically distinct signs. Since existing benchmarks lack ground-truth annotations for sign-level visual similarity, we use sign embedding similarity as a proxy. Once identified, the same target words are used across all methods, ensuring a fair comparison. All target words are selected exclusively from the test split to prevent train–test leakage.

Controlled Comparison and Training Protocol All methods are evaluated under a strictly controlled environment, ensuring that the same target words and substitution positions are used across all models. For each target word, we generate hard negative captions by replacing it with negative candidate words produced by the SAN and the comparative language models. In total, each method generates 10 hard negative captions per sentence, resulting in 40 negatives for each original caption. The fine-grained evaluation assesses how accurately the model retrieves the original caption from among these hard negative captions. During training, we employ multi-word substitutions to improve optimization stability, while fine-grained evaluation uses single-word substitutions to create more challenging minimally perturbed queries. The impact of negative difficulty and substitution granularity is further analyzed in Section 5.

Evaluation Metrics We report Recall at rank k (R@k) for $k \in \{1, 5, 10\}$ and Mean Reciprocal Rank (MRR).

Implementation Details We adopt GFSLT-VLP (Zhou et al., 2023) and CiCo (Cheng et al., 2023) as our base architectures and train all models with the **Cross-Lingual Contrastive Learning (CLCL)** objective (Cheng et al., 2023). All models are trained for 100 epochs using SGD with an initial learning rate of 1×10^{-2} and a cosine learning rate scheduler. We use a batch size of 32 for GFSLT-VLP and 256 for CiCo. For the SAN framework, we set $\alpha = 0.7$ and $\beta = 0.7$ for mining negatives in both training and evaluation, with the balancing weight $\lambda = 0.4$ applied during training. We utilize a trained GFSLT-VLP-based sign language retrieval model for the SAN framework.

Language Models for Negative Word Mining Previous vision-language retrieval studies have demonstrated the effectiveness of using language models to generate hard negative captions for fine-grained visual understanding (Zhang et al., 2024; Momeni et al., 2023; Doveh et al., 2023). Following this line of research, we employ **RoBERTa** (Liu et al., 2019) and **GPT-4o-mini** (OpenAI, 2023) as comparative language models for hard negative generation. We also incorporate **FastText** (Mikolov et al., 2018), motivated by its use in NLA-SLR (Zuo et al., 2023) to enhance discrimination among visually similar signs in sign language recognition.

4.2 Retrieval Results

In Table 1, we compare coarse- and fine-grained retrieval performance under different hard-negative mining strategies.

Unlocking Fine-grained Capabilities with SAN We first investigate the impact of our proposed SAN strategy to validate its effectiveness in enhancing fine-grained discrimination. As shown in

Model	Hard Negatives	Fine-grained (↑)				Coarse-grained (↑)							
		V2T				T2V				V2T			
		R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
Cico	∅	17.9	55.3	79.1	35.0	**69.2**	87.2	**92.2**	**77.3**	**70.1**	**87.7**	**92.9**	**78.2**
	FastText	33.8	**76.8**	92.1	51.8	67.3	86.1	91.6	75.4	67.0	86.0	91.1	75.4
	RoBERTa	25.6	68.2	88.9	43.2	67.1	86.5	91.9	75.7	63.6	84.7	89.4	73.1
	GPT4o-mini	30.7	76.3	**92.7**	49.1	67.5	87.1	91.6	75.9	67.6	86.1	91.3	75.8
	SAN (Ours)	**39.4**	75.4	92.5	**54.4**	68.1	**87.4**	91.7	76.6	67.8	87.4	91.7	76.2
GFSLT-VLP	∅	16.8	53.1	78.0	33.9	67.9	88.4	93.8	77.5	**69.4**	88.7	**93.3**	**77.9**
	FastText	43.8	85.6	**96.7**	61.3	68.6	**90.1**	93.2	77.7	64.3	84.8	89.9	73.4
	RoBERTa	29.5	73.9	95.0	47.9	68.0	89.6	93.9	77.1	65.2	86.7	90.8	74.3
	GPT4o-mini	37.7	85.3	96.4	56.4	**70.9**	89.4	94.1	**79.1**	66.6	**89.1**	92.1	76.2
	SAN (Ours)	**49.1**	**85.9**	94.9	**64.1**	70.2	89.3	**94.4**	78.7	67.4	85.4	90.5	75.5

Table 1: Retrieval performance on coarse-grained and fine-grained evaluation settings on PHOENIX-2014T. **Bold** indicates the best performance.

Table 1, standard retrieval models struggle significantly with fine-grained queries, achieving only 17.9% and 16.8% in V2T R@1 for CiCo and GFSLT-VLP, respectively. However, incorporating our proposed SAN strategy leads to substantial performance gains across both architectures. Specifically, applying SAN to CiCo boosts the fine-grained V2T R@1 from 17.9% to **39.4%**, a remarkable absolute improvement of **+21.5%**. The impact is even more pronounced with GFSLT-VLP, where SAN nearly triples the baseline performance, jumping from 16.8% to **49.1%** (**+32.3%**).

These results suggest that the suboptimal fine-grained performance of existing baselines stems not from insufficient model capacity, but from the lack of exposure to ambiguous instances during training. This highlights the necessity of explicitly learning from confusing negatives: SAN forces the model to resolve these ambiguities, thereby obtaining the discriminative features required to distinguish subtle sign variations.

Comparison with Text-based Mining Strategies We further compare SAN with text-based hard negative mining strategies, including FastText, RoBERTa, and GPT-4o-mini. As illustrated in Table 1, SAN consistently outperforms all text-driven baselines on the fine-grained test set. On the GFSLT-VLP backbone, our method achieves a V2T R@1 of **49.1%**, surpassing the strongest text-based competitor, FastText (43.8%), by a clear margin. Notably, leveraging a state-of-the-art LLM (GPT-4o-mini) yields only 37.7%, falling short of our visual-aware approach by **11.4** percentage points. A similar trend is observed with CiCo, where SAN (39.4%) significantly outperforms GPT-4o-mini (30.7%) and RoBERTa (25.6%). These results underscore a critical insight: *linguistic hardness does not equate to visual hardness* in sign language. While text-based methods generate semantically related negatives (e.g., synonyms), these words often look visually distinct when signed. In contrast, SAN directly mines negatives based on visual similarity, providing much more effective supervision for distinguishing subtle sign variations.

Robustness on Coarse-grained Retrieval Enhancing fine-grained discrimination often entails a trade-off, potentially degrading performance on standard coarse-grained benchmarks due to the distortion of global semantic alignment (Chen et al., 2024). However, our results indicate that SAN offers a significantly more favorable trade-off compared to text-based mining strategies. On the GFSLT-VLP backbone, text-based methods notably degrade coarse-grained V2T performance (e.g., FastText drops R@1 from 69.4% to 64.3%). In contrast, SAN maintains a robust performance of 67.4%, mitigating the degradation. A similar trend is observed with CiCo, where SAN exhibits the smallest performance drop among all mining-based methods. This suggests that visual-aware negatives not only refine detailed visual understanding but can also effectively maintain coarse-grained alignment, whereas linguistic negatives often introduce noise that disrupts standard retrieval.

5 Analysis

Visual Comparison of Mined Negative Words Figure 3 provides a qualitative comparison of the mined negative words in terms of signing motion. In each example, the first row shows the sign corresponding to the original word, the second row shows the negative word mined by SAN (ours), which other methods failed to mine, and the third row shows a negative word mined by text-based methods. Across all cases, SAN consistently mines negatives whose signs exhibit highly similar hand-

(a) Europa (b) Winter

(c) Deutschland (d) Norden

Figure 3: Qualitative comparison of mined hard negative words. For each example, the first row shows the sign of the target word, the second row shows the negative word mined by SAN, and the third row shows a negative mined by text-based methods. SAN retrieves visually similar signs with subtle motion differences, whereas text-based methods select linguistically plausible but visually distinct negatives. The scores indicate the similarity of the sign video features extracted by the pretrained I3D model.

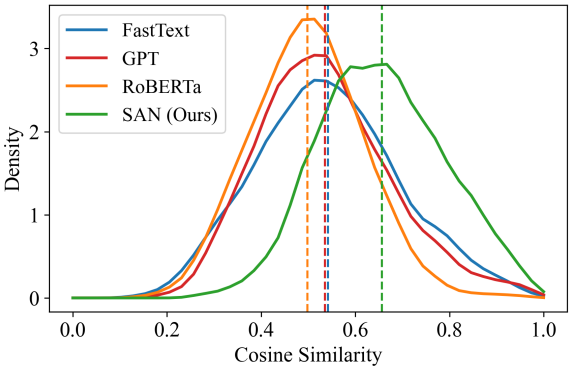


Figure 4: Distribution of cosine similarities between sign prototypes of target and negative words. The distribution of SAN is shifted toward higher similarity values, demonstrating better coverage of visually confusable regions than text-based methods.

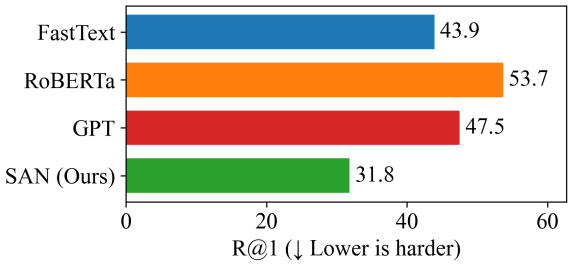


Figure 5: Comparison of the hardness of generated negatives. Lower retrieval performance implies higher difficulty.

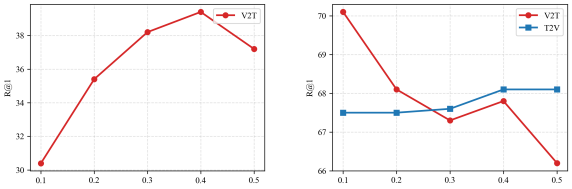
shape, location, and movement to the original sign (e.g., Europe → October, Winter → Cold, Germany → North, North → North Sea), resulting in subtle motion differences that are difficult to distinguish visually. In contrast, negatives mined by text-based methods (e.g., Scandinavia, Spring, Austria, South) often involve clearly different articulations, making them less visually confusable despite being linguistically plausible substitutions. Overall, these examples suggest that SAN can mine true hard negatives—negatives that are highly confusable in sign motion.

Quantitative Analysis of Negative Distribution Mismatch We further analyze the degree of visual confusability of mined negative words. To this end, we establish word-to-sign alignments via the trained CiCo (Cheng et al., 2023) model and generate sign prototypes by averaging the corresponding sign features for each word. Figure 4 presents the cosine similarity distributions between negative words and their corresponding target words for each mining strategy. Notably, negatives mined by SAN exhibit a clear shift toward higher cosine similarity values compared to those mined by FastText, RoBERTa, and GPT-4o-mini. This shift suggests that SAN is more likely to sample negatives from the visually confusable region, thereby mitigating the negative distribution mismatch problem. By contrast, text-based substitutions fail to adequately cover this region, as they prioritize linguistic plausibility over visual similarity.

Difficulty of Generated Hard Negative Captions To validate the intrinsic quality of the mined negatives, we evaluate the retrieval performance on the specific hard-negative sets generated by each method, rather than a merged set. The goal of this experiment is to quantify the discriminative challenge posed by the generated negatives. As illustrated in Figure 5, we observe a substantial disparity in retrieval performance depending on the mining strategy. While the model maintains high accuracy on negatives generated by language models (e.g., RoBERTa: 53.7%, GPT-4o-mini: 47.5%), its performance drops sharply to 31.8% on the SAN-generated set. Crucially, this performance

N_{swap}	FG (↑)	CG (↑)		N_{hard}	FG (↑)	CG (↑)	
	V2T	T2V	V2T		V2T	T2V	V2T
∅	17.9	69.2	70.1	∅	17.9	69.2	70.1
1	38.6	65.0	63.1	3	35.8	66.7	68.2
2	39.4	68.1	67.8	5	39.4	68.1	67.8
3	34.9	68.1	69.9	7	39.7	67.6	67.3

Table 2: Ablation study on the number of swapped words N_{swap} (left) and the number of hard negatives N_{hard} (right). FG and CG denote fine-grained and coarse-grained, respectively.



(a) Fine-grained (b) Coarse-grained

Figure 6: Effect of λ on fine-grained and coarse-grained retrieval performance.

drop does not indicate model failure, but rather confirms the high quality of our hard negatives. Unlike text-based baselines that produce easily distinguishable negatives, SAN successfully mines visually ambiguous instances that effectively challenge the model's discriminative capabilities.

Impact of Number of Swapped Words In Table 2 (Left), we show the impact of N_{swap}, the number of words replaced when generating hard-negative captions. Replacing only a single word yields a substantial improvement in fine-grained retrieval, but also leads to a noticeable drop in coarse-grained performance. This suggests that extremely hard negatives formed by minimal perturbations can over-constrain the model and disrupt the global representation space. As more words are swapped, coarse-grained performance is better preserved, while fine-grained performance gradually decreases, since the negatives become less targeted to specific visually confusable sign units. Overall, these results highlight the importance of balancing fine-grained discrimination and global robustness, and we select $N_{\text{swap}} = 2$ as it achieves the best trade-off between the two.

Impact of Number of Hard Negatives We further examine the effect of N_{hard}, the number of hard-negative captions used per training instance. As shown in Table 2 (Right), increasing N_{hard} consistently improves fine-grained performance, as the model is exposed to a richer set of visually confusing alternatives. However, using too many hard negatives gradually degrades coarse-grained retrieval, indicating a bias toward local discrimination at the expense of global alignment. Conversely, using too few hard negatives provides insufficient fine-grained supervision, leading to suboptimal fine-grained performance. We therefore adopt $N_{\text{hard}} = 5$, which yields near-maximal fine-grained gains while limiting the drop in coarse-grained performance.

Sensitivity to Loss Weighting We analyze the effect of the loss weight λ, which controls the contribution of the hard-negative contrastive objective. As shown in Figure 6, fine-grained retrieval performance consistently improves as λ increases, confirming that prioritizing these negatives effectively sharpens the model's discriminative power. However, overly large λ leads to a noticeable degradation in coarse-grained retrieval, suggesting that excessive emphasis on hard negatives can distort the global video–text embedding structure required for general retrieval. Based on this trade-off, we select $\lambda = 0.4$, which provides strong fine-grained gains while keeping coarse-grained performance largely stable.

6 Conclusion

In this paper, we identify a critical limitation of current sign language retrieval models: their inability to distinguish visually similar signs in fine-grained retrieval scenarios. We show that this limitation does not stem from insufficient model capacity, but from a negative distribution mismatch in contrastive learning, where training supervision fails to reflect the visual ambiguity inherent in sign language. To address this issue, we propose Sign-Aware Hard Negative Mining (SAN), which redefines hard negatives based on visual confusability in the sign embedding space rather than linguistic similarity. Experiments on PHOENIX-2014T demonstrate that SAN substantially improves fine-grained retrieval performance while preserving coarse-grained accuracy, outperforming language-model–based negative mining strategies. These results highlight the importance of shifting negative supervision signals from linguistic similarity to visual similarity of signs, suggesting that visually grounded negative mining is essential for fine-grained sign language retrieval.

Limitations

Our experiments are limited to the PHOENIX-2014T dataset within a single domain, and the generalizability of SAN across other sign languages and datasets remains to be validated. However, we believe that PHOENIX-2014T provides a particularly suitable testbed for studying fine-grained sign language retrieval, as its weather-forecast domain naturally exhibits a dense distribution of visually similar signs (e.g., regional names, numbers, and weather-related terms), creating a concentrated environment for evaluating visual confusability. Furthermore, although we employed POS-filtering to ensure grammatical consistency during fine-grained evaluation, some linguistic noise may still persist in the generated captions. Additionally, the current framework relies on a fixed visual similarity threshold (β), which may not account for the varying degrees of ambiguity across different sign classes; thus, a dynamic or adaptive thresholding mechanism is needed. Lastly, while SAN enhances fine-grained discrimination, it can lead to a slight performance trade-off in coarse-grained retrieval, suggesting the need for balanced optimization strategies to maintain overall retrieval robustness.

Discussion

Phonological features of signs (Sandler, 2012), such as handshape, location, and movement, provide a principled basis for defining sign similarity and offer an alternative avenue for constructing hard negatives. In particular, dictionary-level phonological features can be used to identify *minimal pairs*—sign pairs differing in only a single phonological component—which serve as a proxy for visual hardness without requiring visual data. While this is a promising direction, phonological features are discrete and static, and may not fully capture the continuous visual dynamics of actual signing that embedding-based approaches like SAN can directly model. Moreover, dictionary-based approaches are limited to standardized vocabulary, whereas vision-based mining offers broader coverage of real-world signs. Notably, SAN-mined negatives exhibit phonological similarity (Figure 3), suggesting that the learned embedding space implicitly captures phonological structure. Integrating phonological minimal pairs as complementary signals is a promising direction for future work.

Ethics Statement

For qualitative visualization purposes, we use sign language video examples obtained from the SignDict database[2]. According to the SignDict website, all videos are released under a Creative Commons license. We use these videos solely for academic illustration and analysis, in accordance with the stated licenses.

Acknowledgments

This research was supported by the InnoCORE program of the Ministry of Science and ICT(N10260002) and the Top-Tier AI Global HRD invitation program (RS-2025-25461932) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

Junseok Ahn, Youngjoon Jang, and Joon Son Chung. 2024. Slowfast network for continuous sign language recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 3920–3924. IEEE.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 35–53. Springer.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE.

Aozhu Chen, Hazel Doughty, Xirong Li, and Cees G. M. Snoek. 2024. Beyond coarse-grained matching in video-text retrieval. In *Computer Vision - ACCV 2024 - 17th Asian Conference on Computer Vision, Hanoi, Vietnam, December 8-12, 2024, Proceedings, Part III*, volume 15474 of *Lecture Notes in Computer Science*, pages 25–43. Springer.

[2] https://signdict.org/

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19016–19026. IEEE.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogério Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision & language concepts to vision & language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2657–2668. IEEE.

Amanda Cardoso Duarte, Samuel Albanie, Xavier Giró-i-Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14074–14084. IEEE.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18362–18372. IEEE.

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

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 297–304. JMLR.org.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11067–11076. IEEE.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3901–3920. Association for Computational Linguistics.

Longtao Jiang, Min Wang, Zecheng Li, Yao Fang, Wengang Zhou, and Houqiang Li. 2024. SEDS: semantically enhanced dual-stream encoder for sign language retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 5141–5150. ACM.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3413–3423. Computer Vision Foundation / IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 100. BMVA Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tomás Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in action: Improving verb understanding in video-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15533–15545. IEEE.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching CLIP to count to ten. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3147–3157. IEEE.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024. Tripletclip: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3451–3461. Computer Vision Foundation / IEEE.

Wendy Sandler. 2012. The phonological organization of sign languages. *Lang. Linguistics Compass*, 6(3):162–182.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16857–16866. Computer Vision Foundation / IEEE.

Xuan Wu, Hongxiang Li, Yuanjiang Luo, Xuxin Cheng, Xianwei Zhuang, Meng Cao, and Keren Fu. 2024. Uncertainty-aware sign language video retrieval with probability distribution modeling. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIV*, volume 15102 of *Lecture Notes in Computer Science*, pages 390–408. Springer.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of- rs, and what to do about it? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

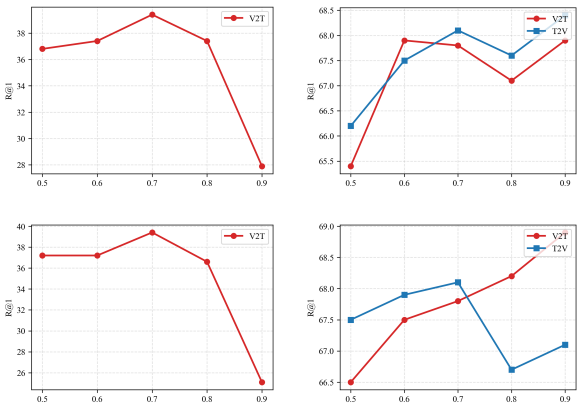
Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13774–13784. IEEE.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. BEST: BERT pre-training for sign language recognition with coupling tokenization. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 3597–3605. AAAI Press.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20814–20824. IEEE.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14890–14900. IEEE.



(a) Fine-grained (b) Coarse-grained

Figure 7: Ablation study on hyperparameters α (top row) and β (bottom row). The left column shows fine-grained retrieval results (V2T), while the right column shows coarse-grained results (V2T and T2V).

A Appendix

A.1 Impact of Threshold α and β in SAN framework

We analyze the effects of two key thresholds in the SAN framework: the reliability threshold α used for sign–word pair mining, and the visual similarity threshold β used for selecting visually confusable negative words. When evaluating the impact of α, we fix β to 0.7, and vice versa, to isolate the effect of each threshold. Figure 7 shows that increasing α progressively filters out noisy sign–word alignments, leading to more stable and reliable supervision. As a result, coarse-grained retrieval performance generally improves with larger α. However, excessively large α reduces fine-grained performance, as fewer sign–word pairs remain eligible for negative mining, thereby limiting the diversity and availability of informative hard negatives. Similarly, the threshold β governs the degree of visual similarity required for negative word selection. Smaller values of β admit visually weak or ambiguous negatives that provide limited fine-grained supervision. In contrast, overly large β restricts the candidate pool to a small set of near-duplicate signs, reducing training diversity and weakening the learning signal. Based on these observations, we set both α and β to 0.7 in all experiments. This choice provides a balanced trade-off between alignment reliability and candidate coverage, yielding consistently strong fine-grained improvements while maintaining stable coarse-grained retrieval performance across models.

A.2 Language-Model-Based Negative Mining Methods

Recent studies have leveraged language models to generate hard negative captions for the vision–language retrieval task (Zhang et al., 2024; Momeni et al., 2023; Doveh et al., 2023). Following this line of work, we compare SAN with several representative language-model-based negative mining strategies, including FastText, RoBERTa, and GPT-4o-mini. For fine-grained evaluation, we fix a set of target words in advance and independently mine negative word candidates for each target word using different strategies. Each method generates exactly ten negative candidates per target word, and these candidates are subsequently used to construct method-specific fine-grained evaluation sets.

A.2.1 FastText

FastText (Joulin et al., 2016) is a static word embedding model that represents words as compositions of character n-grams. We use the pretrained *cc.de.300.bin* model, which is trained on German Common Crawl and Wikipedia corpora. To mine negative candidates, we embed all words in the dataset vocabulary and compute cosine similarities between word embeddings. For each target word, the top-10 most similar words are selected as negative candidates.

A.2.2 RoBERTa

RoBERTa (Liu et al., 2019) is a transformer-based masked language model trained with dynamic masking and large-scale corpora to produce contextualized representations. We employ *XLM-RoBERTa-base*, a multilingual variant trained on 100 languages. Negative candidates are obtained by replacing the target word in the original sentence with a `[MASK]` token and selecting the top-10 words with the highest unmasking probabilities.

A.2.3 GPT-4o-mini

GPT-4o-mini (OpenAI, 2023) is a lightweight autoregressive large language model designed for efficient text generation. We use GPT-4o-mini to generate negative word candidates by prompting the model to suggest ten distinct alternatives for each masked target word, excluding the original word. Unlike FastText and RoBERTa, GPT-4o-mini directly generates candidates through conditional text generation rather than embedding similarity or masked prediction.

The prompt used for negative candidate generation is shown below:

```
For each [MASK] token in the sentence, provide exactly
10 appropriate negative candidate words.
Do NOT provide the original word or any repetitions.
Each candidate must be written in lowercase.
Do NOT include the full sentence itself.
Output only in the following format:

You must use the origin_word exactly as provided.
<origin_word> : <candidate1>,..., <candidate10>
```

Part-of-Speech Control Because SAN and FastText may produce negative candidates with different part-of-speech (POS) tags from the target word, we apply an additional POS filtering step during fine-grained evaluation set construction. Specifically, we use GPT-4o-mini to retain only candidates that share the same POS tag as the target word. This ensures that fine-grained evaluation focuses on visual confusability rather than trivial grammatical inconsistencies.

A.3 Analysis of Mined Negative Words

All sign videos in Figure 9 come from the SignDict database[3]. Visual similarity scores are computed by measuring cosine similarity between sign clip features extracted by the I3D model pre-trained on the BSL-1K (Varol et al., 2021) dataset.

For ***Norden***, SAN primarily mines words associated with northern regions, such as *Nordsee*, *Nordwesten*, and *Nordseeküste*, whereas text-based methods tend to select direction-related words such as *Osten*, *Westen*, and *Süden*. In terms of signing motion, the sign for *Norden* involves raising the hand vertically. The sign for *Nordsee* is visually similar, also exhibiting an upward movement, but differs subtly in the final posture, where the fingers bend by approximately $90°$. By contrast, the sign for *Süden* is visually distinct from *Norden*, characterized by a downward hand movement with a slightly bent wrist.

For ***Winter***, SAN mines words associated with winter-related phenomena, including *kalt*, *Frost*, and *Wind*, while text-based approaches predominantly select season-related words such as *Herbst*, *Sommer*, and *Frühling*. The signs for *Winter* and *Kalt* are nearly identical: both involve a clenched fist, with *Kalt* differing only in the absence of pursed lips and a slightly more inward-oriented fist. In contrast, the sign for *Frühling* is visually dissimilar, involving an upward unfolding motion of the right hand from a closed fist.

For ***Deutschland***, SAN mines words related to northern regions, such as *Norden*, *Nordsee*, and *Nordmeer*, whereas text-based methods select semantically related country names including *Spanien*, *Europa*, and *Österreich*. Both the *Deutschland* and *Norden* signs involve an upward hand movement, although the hand is raised higher in the *Deutschland* sign. By contrast, the sign for *Österreich* differs substantially, as it involves crossing both hands over the chest, briefly clenching the fingers, and then extending them outward.

Across these examples, negative words mined by SAN consistently exhibit higher visual similarity to the target signs than those selected by text-based methods (e.g., *Nordsee* 0.96 vs. *Süden* 0.90; *Kalt* 0.95 vs. *Frühling* 0.72; *Norden* 0.96 vs. *Österreich* 0.78). These observations demonstrate that SAN effectively identifies hard negatives that are visually confusable with the target signs at the gesture level.

Moreover, we further compare the negative words mined by SAN and language-model-based methods in Table 3. The examples show that SAN mines visually similar negatives across both semantically similar and semantically distinct word pairs. For example, SAN captures subtle variations among semantically related words such as *Norden–Nordsee*, as well as visually similar but semantically distinct pairs including *Deutschland–Donnerstag*, *August–Bayern*, and *Europa–Oktober*. These results indicate that SAN effectively mines hard negatives from visually confusable regions, covering both semantically similar and semantically distinct cases that are often missed by language-model-based methods. In contrast, language-model-based methods predominantly select negatives based on linguistic or semantic similarity, which may overlook visually confusable sign pairs. For visual references of the corresponding signs, refer to the SignDict database[4]. Overall, this comparison highlights a clear difference in the types of negatives produced by each approach: SAN emphasizes visual similarity in sign gestures, while language-model-based methods focus on linguistic or semantic similarity.

A.4 Retrieval Example

Figure 8 presents a representative example of fine-grained sign language retrieval on PHOENIX-2014T. In this example, only SAN correctly ranks the target sign video at Rank@1, while baseline methods are confused by visually similar alternatives. The hard negative caption differs from the

[3] https://signdict.org/

(a) *north*

(b) *northwest*

Figure 8: An example of fine-grained retrieval in PHOENIX-2014T where only SAN correctly ranks the target at R@1, while language-model-based methods fail to correctly distinguish visually confusable negatives.
Target caption: "In the north, stormy gusts are also possible."
Hard negative caption: "In the northwest, stormy gusts are also possible."

target caption by a single word (*north* vs. *northwest*), yet the corresponding signs exhibit high visual similarity. This example illustrates how SAN improves fine-grained discrimination by effectively handling visually confusable sign pairs.

Figure 9: Visual examples of the mined hard negative words. Each row shows the sign gestures of the origin word, the word mined by SAN, and the word mined by other methods. The numbers in parentheses indicate visual similarity scores with respect to the origin word.

Table 3: Comparison of mined hard negative words. SAN captures visually confusable negatives across both semantically similar and semantically distinct word pairs, whereas language-model-based methods mainly rely on linguistic similarity.

Word	**Ours**	**Others**
Norden (North)	Nordsee (North Sea)	Osten (East)
	Deutschland (Germany)	Westen (West)
	Nordwesten (Northwest)	Süden (South)
	Nordhälfte (Northern Half)	Südwesten (Southwest)
	Nordosthälfte (Northeastern Half)	Südosten (Southeast)
	Norddeutschland (Northern Germany)	Norddeutschland (Northern Germany)
	Nordseeküste (North Coast)	Küste (Coast)
Winter (Winter)	Kalt (Cold)	Herbst (Autumn)
	Frost (Frost)	Sommer (Summer)
	Wind (Wind)	Frühling (Spring)
	Winterwetter (Winter Weather)	Schnee (Snow)
	Frostfrei (Frost-Free)	Jahreszeit (Season)
	Kälter (Colder)	Hochsommer (Midsummer)
	Gefrierpunkt (Freezing Point)	Frühjahr (Early Spring)
Deutschland (Germany)	Norden (North)	Spanien (Spain)
	Nordsee (North Sea)	Europa (Europe)
	Nordmeer (Northern Sea)	Österreich (Austria)
	Nordosten (Northeast)	England (England)
	Donnerstag (Thursday)	Frankreich (France)
	Norddeutschland (Northern Germany)	Tschechien (Czech Republic)
	Nordhälfte (Northern Half)	Russland (Russia)
August	Oktober (October)	September
	Sonntag (Sunday)	Juli (July)
	Bayern (Bavaria)	November
	September	Februar (February)
	Juli (July)	Juni (June)
Europa (Europe)	Norden (North)	Griechenland (Greece)
	Nordwesten (Northwest)	Schweden (Sweden)
	Oktober (October)	Spanien (Spain)
	Mitteleuropa (Central Europe)	Deutschland (Germany)
	Südosten (Southeast)	Italien (Italy)
	Südwesteuropa (Southwest Europe)	Frankreich (France)
	Deutschland (Germany)	Afrika (Africa)
Nacht (Night)	Neun (9)	Stunde (Hour)
	Nordsee (North Sea)	Woche (Week)
	Deutschland (Germany)	Osterabend (Easter Eve)
	Abend (Evening)	Mitternacht (Midnight)
	Donnerstag (Thursday)	Morgen (Morning)
Donnerstag (Thursday)	Deutschland (Germany)	Montag (Monday)
	Montag (Monday)	Freitag (Friday)
	Samstag (Saturday)	Dienstag (Tuesday)
	Norden (North)	Mittwoch (Wednesday)
	Sonnenschein (Sunshine)	Wochenende (Weekend)