SIGN LANGUAGE QUESTION ANSWERING: A NEW TASK, BENCHMARK, AND BASELINE FOR SIGN LANGUAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in sign language understanding (SLU) have led to remarkable progress in tasks such as continuous SL recognition and SL translation. However, these tasks are designed with predefined objectives, requiring models to learn a fixed mapping from sign videos to glosses or spoken-language sentences. As a result, they provide only a limited assessment of whether a model truly understands the semantic content of SL videos. To address this limitation, **we first propose a new task, Sign Language Question Answering (SLQA)**, which evaluates SL understanding by requiring models to answer natural language questions about SL videos. Unlike previous SLU tasks, SLQA provides a more flexible and comprehensive evaluation framework that assesses multiple reasoning capabilities beyond recognition and translation. To facilitate this task, **we further construct two SignQA benchmarks** based on PHOENIX14T and CSL-Daily by automatically generating question-answer pairs from existing gloss and sentence annotations using carefully designed templates. The resulting datasets cover five complementary question categories, including position reasoning, structural reasoning, visual search, gloss recognition, and translation understanding. **Finally, we propose a simple yet effective baseline model** equipped with a Question-Conditioned Modulated Temporal Downsampling module and an in-domain knowledge transfer strategy, enabling effective knowledge transfer from existing SLU tasks while enhancing question-aware temporal feature modeling. Extensive experiments demonstrate that our baseline consistently outperforms representative vision-language models across all question categories, establishing a strong benchmark for future research on SLQA. **Code and datasets will be made publicly available upon acceptance.**

1 INTRODUCTION

Recent advances in Sign Language(SL) understanding(SLU) have led to significant progress across a wide range of research tasks, including isolated SL recognition (ISLR) (Hu et al., 2021), continuous SL recognition (CSLR) (Min et al., 2021; Gan et al., 2024), gloss-based SL translation (GBSLT) (Chen et al., 2022b), gloss-free SL translation (GFSLT) (Gan et al., 2026a), etc.Specifically, ISLR focuses on classifying an isolated sign into a predefined sign category. Extending this setting, CSLR aims to recognize a continuous SL video and convert it into a sequence of glosses. Building upon CSLR, GBSLT further translates SL videos into spoken sentences with the supervision of gloss annotations. To alleviate the reliance on costly gloss annotations, GFSLT seeks to directly train SLT models to translate SL videos into spoken text without using gloss supervision.

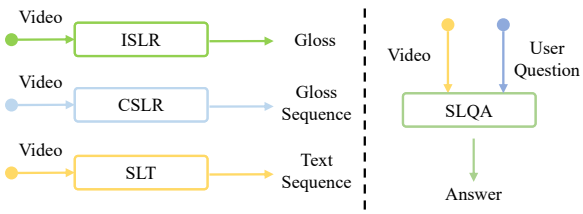


Figure 1: Illustration of existing SLU and proposed SLQA.

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.

However, most of these tasks are task-specific and trained with fixed objectives, where models are optimized to produce predefined outputs (*e.g.*, glosses or text sequences). Although these approaches have advanced the mapping of SL videos to textual representations, they remain limited in terms of flexibility and generalization. In particular, (i) they struggle to accommodate diverse and open-ended user requirements, and (ii) their fixed objectives make it difficult to determine whether a model has genuinely learned to understand the semantic content of SL videos, rather than merely learning task-specific mappings or memorizing the training data.

To address these limitations, inspired by the success of Video Question Answering (VideoQA) (Zhong et al., 2022; Wu et al., 2025; Ataallah et al., 2024; Li et al., 2024), (1) **we propose a new task termed Sign Language Question Answering (SLQA)**. Unlike existing SL tasks that focus on predefined objectives, such as sign classification, gloss recognition, or video-to-text translation, SLQA requires a model to answer diverse natural-language questions based on a SL video. Specifically, the model takes a SL video and a natural language question as input, and generates the corresponding answer. This formulation allows users to query different aspects of the video content, including spatial and temporal relations, gloss sequences, visual details, and sentence-level semantics, thereby providing a more comprehensive evaluation of SLU than conventional task-specific benchmarks. (2) Besides, to evaluate the proposed SLQA task, **we construct two SLQA datasets** based on widely used SL benchmarks, namely PHOENIX14T and CSL-Daily. Specifically, we leverage existing annotations, including gloss-level annotations and sentence-level text annotations, to automatically construct a large number of SLQA pairs via carefully designed question templates. This allows us to generate diverse and controllable QA samples without requiring additional manual annotation. (3) Furthermore, **we propose SLQAM: a baseline model for SLQA**. To enable effective learning with limited SLQA training data, we introduce two key strategies. First, to better enable the model to attend to question-relevant video regions, we design a **Question-Conditioned Modulated Temporal Downsampling (QCMTD)** module, which explicitly aligns the question semantics with temporal sign video representations, allowing the model to focus on question-relevant spatio-temporal features for answer prediction. Second, we propose domain knowledge transfer, where the model is pre-trained on related SL tasks, including CSLR and SLT, and subsequently fine-tuned on the SLQA datasets. We empirically observe that training the model from scratch on SLQA leads to suboptimal performance due to limited data scale, while pre-training significantly improves generalization. Our main contributions are summarized as follows:

- We propose a new SLU task: **SL Question Answering (SLQA)**. Unlike existing SLU tasks with fixed objectives, SLQA focuses on understanding SL videos by answering natural language questions related to the video content.
- We further propose a baseline model for SLQA. To effectively model the relationship between SL videos and questions, we introduce two key components: (1) **QCMTD**, which replaces fixed temporal downsampling and dynamically modulates informative temporal segments conditioned on the input question, thereby enhancing question-relevant video representations; (2) **In-domain knowledge transfer training**, which leverages three stage training pipeline to transfer domain-specific knowledge into SLQA learning.
- We construct two SLQA datasets and establish comprehensive benchmarks covering five complementary question categories. Extensive experiments demonstrate that our approach consistently outperforms existing models across these categories.

2 RELATED

Sign Language Understanding. Existing research in SLU has primarily focused on a set of task-specific objectives, including isolated SL recognition (ISLR), continuous SL recognition (CSLR), and SL translation (SLT). ISLR classifies an isolated sign into a predefined category (Zhou et al., 2025; Zuo et al., 2023), while CSLR recognizes continuous SL videos as sequences of glosses. Building upon CSLR, SLT aims to translate SL videos into spoken-language sentences. GBSLT typically leverages gloss annotations as intermediate supervision, whereas GFSLT seeks to directly learn the alignment between sign videos and textual semantics without gloss supervision (Zhou et al., 2023; Guo et al., 2025). Despite their advances, these tasks are fundamentally defined by fixed objectives: a model is trained to map an SL video to a predefined type of output. Such task-specific formulations provide limited flexibility for evaluating the diverse capabilities required to understand

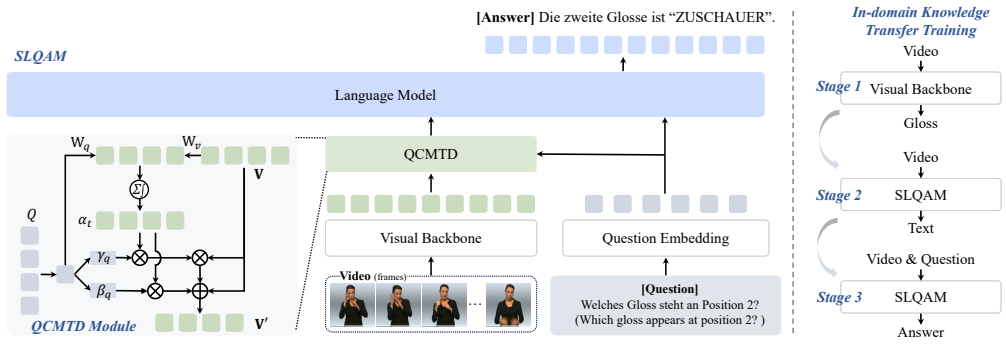


Figure 2: Overall architecture of the proposed SLQAM. The framework consists of a QCMTD module and an in-domain knowledge transfer module.

an SL video. In particular, they do not allow a model to flexibly answer different questions about the same video, making it difficult to assess whether the model can selectively retrieve and reason over different aspects of the visual content.

Recent studies have begun to incorporate questions into SLU. Representative works, such as SSL-SSAW (Liu et al., 2026c) and QD-SLT (Gao et al., 2024), introduce questions as additional conditions for SLT. Although different questions may guide the model to focus on different semantic aspects of the same SL video, the ultimate objective remains translation generation. Thus, the question serves as an auxiliary condition for generation rather than the target of video-grounded reasoning itself. SLU-2K (Testa et al., 2026) further introduces question-answer pairs for SL understanding. However, its questions and answers are confined to the textual translation, and the task primarily evaluates comprehension of the translated sentence rather than understanding the visual content of the SL video.

In contrast, we formulate SL Question Answering (SLQA) as a bora SLU task in which questions are directly grounded in SL videos. Given an SL video and a natural-language question, the model must answer the question by retrieving and reasoning over the relevant visual content, rather than simply generating a fixed-form output or reasoning over a pre-existing translation. Our benchmark covers five complementary question categories, including *Position Reasoning*, *Structural Reasoning*, *Visual Search*, *Gloss Recognition*, and *Translation Understanding*, which collectively evaluate different levels of SL understanding, from fine-grained visual retrieval and sequential reasoning to gloss recognition and sentence-level semantic understanding. Therefore, SLQA provides a more flexible evaluation paradigm for assessing whether a model can genuinely understand the content conveyed by an SL video.

Image Question Answering. Visual Question Answering (VQA) lies at the intersection of computer vision and natural language processing, requiring models to understand visual content and perform reasoning over natural language queries. According to the modality of visual input, VQA can be broadly categorized into Image Question Answering (Image QA) and Video QA. Early Image QA methods primarily focused on learning joint visual–textual representations through attention mechanisms and multimodal fusion modules. Benchmark datasets such as VQA (Antol et al., 2015) and CLEVR (Johnson et al., 2017) have driven the development of bilinear pooling and co-attention networks for modeling fine-grained spatial–language interactions. With the emergence of Transformer architectures, Vision-Language Pretraining (VLP) models such as ViLBERT (Lu et al., 2019), BLIP (Li et al., 2022), and Flamingo (Alayrac et al., 2022) further improved performance by leveraging large-scale image–text pairs to learn cross-modal alignment. More recently, Vision-Language Models (VLMs) and Multimodal LLM, including LLaVA (Liu et al., 2023) and GPT-4V, have shifted the paradigm toward instruction-tuned LLM-based reasoning, enabling open-ended and zero-shot visual question answering with strong generalization capability.

Video Question Answering. Compared with Image QA, Video QA is inherently more challenging due to the additional temporal dimension, which requires reasoning over motion patterns, event ordering, and long-range temporal dependencies. Early approaches extended image-based architectures using RNNs, 3D CNNs, and temporal attention mechanisms to aggregate frame-level features, evaluated on benchmarks such as MSRVTT-QA (Xu et al., 2017) and ActivityNet-QA (Yu et al.,

2019). To better model complex temporal structures, later methods introduced graph-based models and spatio-temporal Transformers, treating frames or object-level entities as nodes to capture cross-frame dependencies. In the current era of foundation models, adapting large language models to video understanding has become the dominant paradigm. Modern video-LLMs (*e.g.*, Video-LLaVA (Lin et al., 2024), Video-ChatGPT (Maaz et al., 2024), and LLaMA-VID (Li et al., 2024)) typically rely on visual encoders combined with token compression or resampling modules to reduce the large number of video tokens while preserving essential temporal semantics.

Despite the remarkable progress of Video QA, existing methods are developed for natural videos and evaluated on open-domain benchmarks. SL videos, however, exhibit unique characteristics, including fine-grained hand articulations, non-manual markers, and strong temporal dependencies, which require substantially different visual representations and reasoning capabilities. Furthermore, unlike conventional Video QA that focuses on objects, events, and scene understanding, the proposed SLQA task requires reasoning over the linguistic content encoded in sign videos, including gloss recognition, structural reasoning, visual search, and translation understanding. Therefore, existing Video QA methods cannot be directly applied to this new setting.

3 METHOD

Overview. In this paper, we propose a new task, termed Sign Language Question Answering (SLQA). Specifically, given a SL video $V = \{f_t\}_{t=1}^{T}$ with T frames and a natural language question Q, the objective of SLQA is to generate the corresponding answer A conditioned on both the video and the question, i.e., $P(A \mid V, Q)$. To facilitate research on this task, we first construct two SLQA datasets based on two widely used public SL datasets, PHOENIX14T and CSL-Daily. By leveraging the existing gloss and text annotations together with carefully designed question templates, we automatically generate diverse question-answer pairs covering five representative categories: *Position Reasoning*, *Structural Reasoning*, *Visual Search*, *Gloss Recognition*, and *Translation Understanding*. Furthermore, we develop a baseline model for SLQA. To effectively bridge SLU and question answering, we introduce two key designs: (1) a *Question-Conditioned Modulated Temporal Downsampling* module that dynamically selects question-relevant video frames instead of using conventional fixed temporal downsampling, and (2) an *in-domain knowledge transfer* strategy that transfers knowledge learned from CSLR and SLT to the SLQA task. Finally, we establish a comprehensive benchmark on the proposed datasets and conduct extensive experiments to evaluate representative methods and the proposed baseline, providing a strong foundation for future research on SLQA.

3.1 DATASET CONSTRUCTION.

To facilitate the SLQA task, we construct two SLQA datasets based on two widely used SLT datasets, PHOENIX14T and CSL-Daily. Instead of collecting free-form question-answer pairs through manual annotation, we automatically generate QA pairs by leveraging the existing gloss annotations and spoken-language translations. Specifically, we design a collection of question templates that transform the original annotations into diverse question-answer pairs. This template-based construction strategy ensures that the generated answers are unambiguous, automatically verifiable, and scalable, while requiring no additional human annotation.

Question Categories. To comprehensively evaluate different aspects of SL understanding, we design five complementary question categories. The proposed benchmark covers both low-level visual reasoning (M1–M3) and high-level linguistic understanding (M4–M5), providing a comprehensive evaluation of sign language understanding from multiple perspectives.

- **M1 – Position Reasoning.** Given a position in the gloss sequence (e.g., the first, last, middle, or N-th position), the model is required to identify the corresponding gloss.
- **M2 – Structural Reasoning.** The task requires understanding the structural relationships within the gloss sequence, such as identifying neighboring glosses or the gloss located between two specified signs.
- **M3 – Visual Search.** The model is required to localize specific visual content in a SL video, including retrieving the position of a target gloss or identifying non-manual markers (e.g., interrogative facial expressions). The latter is instantiated only for CSL-Daily since PHOENIX14T contains only declarative utterances.

- **M4 – Gloss Recognition.** The model answers questions that require recognizing the gloss sequence corresponding to the input SL video.
- **M5 – Translation Understanding.** The model answers questions requiring the spoken-language translation of the input SL video.

Generation Procedure. For each question category, we design a set of manually written question templates with multiple paraphrases to improve linguistic diversity. Given an SL video, we first sample a template from the corresponding category and instantiate its placeholders (e.g., glosses, positions, or gloss pairs) using the GT gloss annotations. The answer is then deterministically derived from the annotations according to the question type, ensuring that every generated QA pair is accurate and automatically verifiable. To prevent models from exploiting simple answer priors, we additionally generate a portion of negative examples. Specifically, some questions intentionally refer to non-existent glosses, out-of-range positions, or ambiguous references caused by repeated glosses, where the correct answers are "not found," "out of range," or "ambiguous," respectively. This design encourages genuine reasoning over the sign sequence instead of memorizing frequent answer patterns. Overall, we design 30 question templates across the five question categories, with multiple paraphrases for each template in both German (PHOENIX14T) and Chinese (CSL-Daily). All random sampling is performed with fixed seeds to ensure exact reproducibility. Representative examples are provided in the Appendix B.

3.2 Baseline Model

Inspired by previous SLU models (Min et al., 2021; Gan et al., 2026d), as shown in Figure 2, our model for SLQA consists of three main components: a visual backbone, a QCMTD module for temporal downsampling, and a language model for answer generation. Specifically, given a sign language video $V = \{f_t\}_{t=1}^{T}$, the visual encoder first extracts frame-level features as $E_v = \mathcal{VE}(V)$, where $E_v \in \mathbb{R}^{T\times D}$. Then, we apply the QCMTD module to integrate question information and perform adaptive temporal downsampling: $\tilde{\mathbf{E}}_v = \mathcal{QCMTD}(E_v, Q)$, where $\tilde{\mathbf{E}}_v \in \mathbb{R}^{T'\times D}$ and $T' < T$. Finally, the downsampled question-aware video representation is fed into a language model together with the question to generate the final answer: $A = \mathcal{LM}(\tilde{\mathbf{E}}_v, Q)$.

Question-Conditioned Modulated Temporal Downsampling. Most existing SLT pipelines (Gan et al., 2026d; Lin et al., 2023) employ a temporal 1D convolution to downsample video features before translating the whole sentence. This design is effective for SLT, since the target text is usually semantically aligned with the entire video. However, in SLQA, the question often refers only to a specific temporal segment of the video. Directly applying question-agnostic temporal downsampling may suppress the question-relevant frames. To address this issue, we propose a *QCMTD* module, which modulates video features with the question representation before each temporal pooling operation.

Specifically, as shown in Figure 2, given the input SL feature sequence $\mathbf{E}_v = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T] \in \mathbb{R}^{T\times D}$ and the question representation $\mathbf{Q} \in \mathbb{R}^{L\times D}$, we first obtain a global question embedding $\mathbf{q}$ by average pooling over all question tokens. The visual features and the global question embedding are then projected into a shared hidden space:

$$\mathbf{h}_t = \tanh(\mathbf{W}_v\mathbf{e}_t + \mathbf{W}_q\mathbf{q}),$$

where $\mathbf{W}_v$ and $\mathbf{W}_q$ denotes learnable projections. The temporal relevance weight is computed as:

$$\alpha_t = \sigma(\mathbf{w}_a^\top \mathbf{h}_t),$$

where $\sigma(\cdot)$ denotes the sigmoid function. The relevance score α_t indicates how important the t-th frame is with respect to the input question. Instead of directly pooling the original visual features, we use the question feature to generate modulation parameters:

$$\boldsymbol{\gamma}_q = \mathbf{W}_\gamma\mathbf{q}, \qquad \boldsymbol{\beta}_q = \mathbf{W}_\beta\mathbf{q},$$

where $\boldsymbol{\gamma}_q, \boldsymbol{\beta}_q \in \mathbb{R}^D$. The visual feature at each timestep is then modulated as:

$$\tilde{\mathbf{e}}_t = \mathbf{e}_t \odot (1 + \alpha_t\boldsymbol{\gamma}_q) + \alpha_t\boldsymbol{\beta}_q,$$

where $\odot$ denotes element-wise multiplication. This formulation allows the model to enhance or suppress different visual channels according to the question, while the temporal relevance weight

α_t ensures that the modulation is stronger on question-relevant frames. After modulation, temporal downsampling is performed on the question-aware features:

$$\tilde{\mathbf{E}} = \mathrm{Pool}\left(\mathrm{Conv1D}(\tilde{\mathbf{E}})\right),$$

where $\tilde{\mathbf{E}} = [\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_{T'}]$ denotes the modulated video features. In practice, we insert the proposed modulation before each temporal max-pooling layer in the original temporal convolution block. Therefore, the module preserves the efficiency of standard temporal downsampling while making the downsampling process adaptive to the question. This encourages the network to retain video segments that are more informative for answering the given question.

In-domain Knowledge Transfer Training. Training an SLQA model directly on the proposed SLQA datasets often leads to unsatisfactory performance due to the limited amount of training data and the complexity of question-driven reasoning. To address this issue, we propose an *in-domain knowledge transfer* strategy that progressively transfers knowledge from existing sign language understanding tasks to SLQA. Specifically, the training pipeline consists of three stages: (1) visual backbone pre-training via CSLR, (2) semantic knowledge transfer through SLT, and (3) task-specific fine-tuning on SLQA.

Stage 1: CSLR pre-training. Following previous sign language understanding works, we first pre-train the visual encoder $\mathcal{VE}$ on the CSLR task. Given a sign language video f and its corresponding gloss sequence g, the visual encoder is optimized using the Connectionist Temporal Classification (CTC) objective to learn temporally aligned sign representations:

$$\Theta^*_{\mathcal{VE},w} = \arg\min_{\Theta_{\mathcal{VE},w}} \mathbb{E}_{(f,g)\sim\mathcal{D}} \left[\mathcal{L}_{\mathrm{CTC}}\left(\mathcal{VE}(f)\cdot w, g\right)\right], \tag{1}$$

where $\mathcal{L}_{\mathrm{CTC}}$ denotes the CTC loss, w is the gloss classifier, and $\mathcal{D}$ is the CSLR training dataset.

Stage 2: SLT Training. After obtaining a well-trained visual encoder, we use its extracted visual representations as input to a language model $\mathcal{TR}$ (*i.e.*, mT5) for SLT. The language model is optimized using the standard cross-entropy loss:

$$\Theta^*_{\mathcal{TR}} = \arg\min_{\Theta_{\mathcal{TR}}} \mathbb{E}_{(v,t)\sim\mathcal{D}} \left[-\log p_{\Theta_{\mathcal{TR}}}\left(t \mid \mathcal{TR}(v)\right)\right], \tag{2}$$

where $v = \mathcal{VE}(f)$ denotes the visual representation extracted by the pre-trained visual encoder.

Stage 3: SLQA fine-tuning. Finally, the well-trained SLT model is fine-tuned on the proposed SLQA datasets using a token-weighted cross-entropy loss. Since answers are generated from predefined templates, template tokens contribute little to SLU. Therefore, we assign smaller weights w_i to template tokens and larger weights to answer-specific tokens, encouraging the model to focus on predicting the semantic content rather than memorizing fixed answer patterns.

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{(v,q,a)\sim\mathcal{D}} \left[-\sum_{i=1}^{|a|} w_i \log p_\Theta\left(a_i \mid a_{<i}, v, q\right)\right] \tag{3}$$

4 Experiments

Implementation Details. For our baseline model, we adopt ResNet-18 as the backbone for extracting SL features, followed by the proposed QCMTD module to aggregate and compress temporal representations. For answer generation, we employ mT5 as the language decoder. The entire model is implemented in PyTorch 2.6 and trained on two NVIDIA V100 GPUs using mixed-precision training with FP16. During training, we assign different weights to template and answer tokens in the answer-generation loss. Specifically, the weights for the template tokens and answer tokens are set to 0.02 and 0.98, respectively. During the SLT training stage, $\boldsymbol{\gamma}_q, \boldsymbol{\beta}_q$ in QCMTD are set to 0.

Evaluation Metrics. Since SLQA covers multiple question categories, we adopt different evaluation metrics according to the answer format. For **Gloss Recognition** questions, where the answers are gloss sequences, we report the WER, following the standard evaluation protocol for CSLR. For **Translation Understanding** questions, whose answers are spoken-language sentences, we adopt ROUGE-L F1 (Lin, 2004) and BLEU-1/2/3/4 (Papineni et al., 2002), following common practice

Table 1: Performance with different downsampling modules on Phoenix14T-QA. For gloss recognition, we report WER. For structured QA, we report F1. For translation understanding, RL denotes ROUGE-L F1 and B1–B4 denote BLEU-1 to BLEU-4.

Method	Components			Gloss Recognition		Structured QA		Translation Understanding (Dev) ↑				Translation Understanding (Test) ↑			
	α_t	γ_q	β_q	WER (Dev)↓	WER (Test)↓	F1 (Dev)↑	F1 (Test)↑	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B4	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B4
TemConv (Min et al., 2021)				33.22	34.21	50.66	49.34	53.62	43.65	35.78	16.02	53.03	43.02	35.87	12.58
Q-Former (Li et al., 2023)				34.81	36.92	48.57	48.86	46.46	40.62	36.74	16.46	46.84	37.84	36.25	13.45
TemConv + Q-FiLM		✓	✓	33.12	35.04	49.48	49.10	54.19	42.96	36.41	16.74	54.30	42.14	35.47	14.01
TemConv + Temporal Gate	✓			32.07	33.16	50.16	50.00	54.62	42.17	36.75	16.94	54.16	43.07	36.05	14.95
QCMTD	✓	✓	✓	31.67	33.58	51.61	50.57	55.14	44.91	37.93	17.33	55.50	44.85	36.82	15.58

Table 2: Overall SLQA performance on Phoenix14T-QA with different downsampling modules.

Method	Components			Overall Dev ↑					Overall Test ↑				
	α_t	γ_q	β_q	RL	B1	B4	BLEURT	CIDEr	RL	B1	B4	BLEURT	CIDEr
TemConv (Min et al., 2021)				52.42	54.24	32.64	54.67	2.742	52.53	54.53	33.13	54.53	2.642
Q-Former (Li et al., 2023)				53.81	54.47	33.58	55.04	2.843	53.00	53.61	32.41	54.86	2.722
TemConv + Q-FiLM		✓	✓	54.21	55.13	33.89	55.46	2.801	53.74	54.68	33.54	55.36	2.642
TemConv + Temporal Gate	✓			54.63	56.01	34.05	55.72	2.716	53.14	54.16	33.18	55.04	2.648
QCMTD	✓	✓	✓	55.16	57.05	35.30	56.92	2.923	54.16	56.25	34.12	56.21	2.768

Table 3: Effect of the proposed in-domain knowledge transfer strategy on Phoenix14T-QA. Stage 1 denotes CSLR pre-training, while Stage 2 introduces SLT training before fine-tuning on SLQA.

Training Strategy	Gloss Recognition		Structured QA		Translation Understanding (Dev) ↑				Translation Understanding (Test) ↑			
	WER↓ (Dev)	WER↓ (Test)	F1↑ (Dev)	F1↑ (Test)	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B4	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B4
End-to-End	100.00	100.00	40.84	40.65	36.59	20.90	13.52	3.87	36.59	20.38	13.58	3.74
+ Stage 1 (CSLR)	34.74	35.62	50.58	47.90	27.75	25.73	14.97	6.19	27.45	24.90	14.74	6.13
+ Stage 2 (SLT)	31.67	33.58	51.61	50.57	55.14	44.91	37.93	17.33	55.50	44.85	36.82	15.58

in SLT (Gan et al., 2021). For the remaining question categories, including **Position Reasoning**, **Structural Reasoning**, and **Visual Search**, where the answers are short spans or discrete tokens, we report the F1-score to measure answer correctness. In addition, to comprehensively evaluate the semantic quality of generated answers across all question categories **(overall performance)**, we report ROUGE, BLEU, BLEURT and CIDEr. Together, these metrics provide a comprehensive evaluation of both the lexical accuracy and semantic correctness of the generated answers.

5 Ablation Study

To better show the contribution of each component, we conduct ablation studies on the PHOENIX14T-QA dataset.

Ablation on Temporal Downsampling. We first compare QCMTD with two representative temporal aggregation strategies: the temporal 1D convolution (*TemConv*) widely used in CSLR and SLT, and the Q-Former commonly adopted in vision-language models. We further introduce two partial variants to disentangle the contributions of temporal relevance estimation and question-conditioned channel modulation. *TemConv + Q-FiLM* removes the temporal relevance gate α_t and applies the question-conditioned scale γ_q and bias β_q uniformly to all temporal features. In contrast, *TemConv + Temporal Gate* retains the question-dependent relevance score α_t to reweight temporal features, but removes the channel-wise scale and bias. As shown in Tables 1 and 2, the Q-Former performs comparably to TemConv but does not consistently improve over this simple baseline. Both Q-FiLM and temporal gating improve upon TemConv, indicating that question-conditioned channel modulation and temporal relevance estimation are individually beneficial. Combining the two components in QCMTD yields the strongest overall performance, demonstrating their complementary effects. These results suggest that modulating temporal features according to the input question before fixed temporal downsampling helps preserve information relevant to answer generation.

Ablation on In-domain knowledge transfer training. Tables 3 and 4 evaluate the effectiveness of the proposed in-domain knowledge transfer strategy. We compare direct end-to-end training with two successive stages of in-domain pre-training: CSLR pre-training (Stage 1) and subsequent SLT pre-training (Stage 2). Directly optimizing the model on the SLQA objectives leads to substantially inferior performance. This is mainly because SLQA simultaneously requires fine-grained sign perception, temporal alignment, semantic understanding, and language generation, while the available QA data are insufficient to learn all these capabilities from scratch. As a result, the model strug-

Table 4: Overall SLQA performance under different training strategies on Phoenix14T-QA.

Training Strategy	Overall Dev ↑					Overall Test ↑				
	RL	B1	B4	BLEURT	CIDEr	RL	B1	B4	BLEURT	CIDEr
End-to-End	42.38	29.86	14.06	43.52	1.315	42.42	30.93	14.37	43.53	1.236
+ Stage 1 (CSLR)	47.85	46.03	27.86	50.66	2.564	46.61	45.49	27.17	49.90	2.412
+ Stage 2 (SLT)	55.16	57.05	35.30	56.92	2.923	54.16	56.25	34.12	56.21	2.768

Table 5: Comparison on Phenix14T-QA dataset.

Model	Gloss Recognition		Structured QA		Translation Understanding (Dev)↑						Translation Understanding (Test)↑					
	WER↓		F1↑													
	Dev	Test	Dev	Test	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B2	B3	B4	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B2	B3	B4
VideoLLaMA3-2B	100.00	100.00	37.91	37.62	32.78	12.04	13.10	5.08	2.06	1.03	32.66	11.96	12.74	5.31	2.30	1.03
Qwen3-VL-2B-Instruct	93.73	94.89	42.75	42.92	36.46	22.18	23.02	13.28	8.94	6.70	36.55	20.93	23.68	13.76	9.30	7.02
InternVL3-2B	82.42	82.67	47.74	47.38	40.48	25.58	26.32	16.49	11.68	9.01	39.61	25.60	26.6	17.18	12.36	9.59
GFSLT-VLP (Zhou et al., 2023)	46.63	45.65	47.65	46.17	45.12	40.42	30.14	22.46	17.48	10.01	44.92	40.35	30.94	22.31	16.41	10.14
MMTLB (Chen et al., 2022a)	40.68	40.24	48.65	47.52	47.35	40.69	30.65	24.65	17.45	10.98	46.13	40.48	30.64	23.34	16.14	10.72
MixSignGraph (Gan et al., 2026d)	38.16	39.41	49.46	49.75	49.84	42.36	32.47	25.69	19.64	13.54	48.86	42.40	33.47	23.86	18.37	14.04
Sign2Text2Answer	32.34	35.64	50.34	49.34	53.31	43.11	35.21	25.53	20.34	15.65	54.76	43.64	35.42	24.53	19.55	14.62
SLQAM (Sign2Answer)	31.67	33.58	51.61	50.57	55.14	44.91	37.93	27.61	21.39	17.33	55.50	44.85	36.82	25.80	19.53	15.58

Table 6: Comparison of overall performance on Phenix14T-QA datasets.

Model	Overall Dev ↑							Overall Test↑						
	RL	B1	B2	B3	B4	BLEURT	CIDEr	RL	B1	B2	B3	B4	BLEURT	CIDEr
VideoLLaMA3-2B	44.67	43.33	31.68	25.61	20.89	37.24	1.278	44.69	43.19	31.67	25.49	20.69	37.45	1.252
Qwen3-VL-2B-Instruct	47.71	46.72	34.49	28.12	23.39	42.01	1.449	47.70	47.87	35.43	28.85	23.93	42.02	1.410
InternVL3-2B	50.85	50.39	37.85	31.04	26.01	45.15	1.746	50.93	51.08	38.58	31.81	26.83	45.45	1.736
GFSLT-VLP (Zhou et al., 2023)	50.65	50.61	39.65	35.14	30.41	49.31	1.972	49.16	49.64	38.64	33.64	29.64	48.56	1.904
MMTLB (Chen et al., 2022a)	52.76	53.47	42.56	38.65	32.65	53.64	2.402	52.86	52.98	41.41	37.68	31.03	52.71	2.341
MixSignGraph (Gan et al., 2026d)	53.78	55.67	45.74	38.98	33.05	54.14	2.536	53.36	56.74	44.46	38.67	34.01	55.61	2.501
Sign2Text2Answer	53.34	55.53	45.76	39.53	34.04	54.42	2.433	53.54	54.96	44.01	38.12	33.24	54.54	2.343
SLQAM (Sign2Answer)	55.16	57.05	46.60	40.16	35.30	56.92	2.923	54.16	56.25	45.58	39.02	34.12	56.21	2.768

gles to establish reliable visual representations and cross-modal semantic alignment through the QA objective alone. In contrast, in-domain knowledge transfer provides a progressive learning path. CSLR pre-training first equips the visual encoder with sign-specific representations and temporal alignment through gloss-level supervision, which is particularly beneficial for recognition and fine-grained reasoning. Building upon these visual representations, SLT pre-training further introduces sentence-level semantic alignment and language generation capabilities. This enables the model to bridge visual sign representations with natural language before being fine-tuned for the more diverse SLQA objectives. Consequently, the two stages provide complementary benefits: CSLR establishes the visual and temporal foundation, while SLT further develops cross-modal semantic understanding and generation ability. The consistent improvements across the different SLQA categories and overall evaluation metrics demonstrate the effectiveness of this progressive in-domain knowledge transfer strategy.

6 COMPARISONS

Since SLQA is a newly proposed task, no dedicated baseline is directly available. We therefore construct representative baselines from three complementary categories to provide a comprehensive comparison. (1) **General-purpose Video Understanding Models.** We first consider general-purpose video understanding models that have demonstrated strong capabilities in video-language reasoning. Specifically, we select VideoLLaMA3-2B (Zhang et al., 2025), Qwen3-VL-2B-Instruct (Bai et al., 2025), and InternVL3-2B (Wang et al., 2024). We fine-tune these models on the proposed SLQA datasets to adapt them to sign language question answering. (2) **Task-specific Sign Language Understanding Models.** We further compare with representative SLT models. Since these models are originally designed to generate spoken-language translations rather than answer questions, we adapt their output heads and fine-tune them on the proposed SLQA datasets, enabling a direct comparison under the SLQA setting. (3) **Cascaded Sign2Text2Answer Pipeline.** A natural alternative to our end-to-end SLQAM is a cascaded pipeline that first converts the input sign video into textual representations and subsequently performs question answering. To evaluate this paradigm, we construct a two-stage baseline based on our trained gloss-based SLT model. Specifically, the GBSLT model first predicts both the gloss sequence and the spoken-language translation of the input video. The predicted gloss sequence and translation, together with the input question, are then fed into a separately trained mT5 model used exclusively for question answering. We denote this cascaded baseline as **Sign2Text2Answer**, in contrast to our end-to-end **Sign2Answer** approach, which directly generates the answer from the sign video and question.

Table 7: Comparison on CSL-daily-QA dataset.

Model	Gloss Recognition		Structured QA		Translation Understanding (Dev)						Translation Understanding (Test)					
	WER↓		F1↑													
	Dev	Test	Dev	Test	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B2	B3	B4	$\text{BLEURT}_{\text{SLT}}$	RL	B1	B2	B3	B4
VideoLLaMA3-2B	100.00	100.00	31.78	31.28	20.03	13.73	12.06	2.56	0.56	0.23	19.97	13.42	11.78	2.57	0.65	0.27
Qwen3-VL-2B-Instruct	100.00	100.00	33.34	33.37	25.13	14.97	15.46	4.30	1.64	0.77	24.59	14.69	15.08	4.15	1.31	0.60
InternVL3-2B	90.63	90.24	45.27	44.70	31.34	21.93	22.55	10.43	5.74	3.63	32.42	22.56	23.23	11.25	6.42	4.09
Sign2GPT (Wong et al., 2024)	46.53	47.85	50.45	56.34	53.61	50.17	50.36	33.18	23.18	19.78	52.87	51.60	50.74	32.70	22.85	18.56
MMTLB (Chen et al., 2022a)	40.78	40.12	55.38	57.68	55.93	51.63	50.89	36.89	27.56	20.96	55.47	51.06	50.26	36.13	27.71	20.74
MixSignGraph (Gan et al., 2026d)	38.36	39.45	58.65	59.16	55.48	51.92	50.02	36.02	27.98	20.65	56.89	51.41	50.41	36.16	27.15	20.76
Sign2Text2Answer	36.74	35.65	60.54	60.53	56.64	52.42	50.53	37.51	27.64	21.51	55.80	52.64	52.56	37.53	28.97	21.64
SLQAM (Sign2Answer)	34.19	34.01	62.56	62.71	58.00	52.99	52.99	39.71	30.43	23.95	57.87	53.16	52.56	39.43	30.30	23.84

Table 8: Comparison of overall performance on CSL-daily-QA dataset.

Model	Overall Dev							Overall Test						
	RL	B1	B2	B3	B4	BLEURT	CIDEr	RL	B1	B2	B3	B4	BLEURT	CIDEr
VideoLLaMA3-2B	60.50	59.01	52.79	47.58	43.68	38.46	1.753	60.51	58.83	52.64	47.4	43.48	38.24	1.735
Qwen3-VL-2B-Instruct	60.13	60.97	54.22	48.70	44.62	38.76	1.715	60.24	60.83	54.14	48.65	44.57	38.32	1.699
InternVL3-2B	64.33	64.53	57.84	52.34	48.21	44.66	2.311	64.27	64.46	57.69	52.14	47.95	44.31	2.277
Sign2GPT (Wong et al., 2024)	65.40	65.36	62.35	59.62	54.36	46.16	2.672	64.59	66.14	61.73	58.07	55.31	45.20	2.947
MMTLB (Chen et al., 2022a)	69.21	70.71	67.12	61.13	57.14	56.78	3.890	69.13	70.06	67.32	61.48	56.16	56.14	3.604
MixSignGraph (Gan et al., 2026d)	68.30	70.41	66.98	61.89	58.35	57.49	3.465	68.64	70.07	67.16	60.34	57.47	57.06	3.750
Sign2Text2Answer	70.61	70.66	66.12	60.13	56.93	54.24	3.647	70.75	70.64	66.03	60.63	55.03	54.04	3.754
SLQAM (Sign2Answer)	73.39	75.31	68.93	63.60	59.16	60.41	4.179	73.52	75.32	68.92	63.60	59.18	60.23	4.193

Evaluation of SLQA Performance on PHOENIX14T-QA. Tables 5 and 6 compare SLQAM with general-purpose VLMs and the cascaded Sign2Text2Answer baseline on PHOENIX14T-QA. General-purpose VLMs perform poorly, particularly on gloss recognition and structured QA, indicating that general video-language pre-training is insufficient to capture the fine-grained hand movements and temporal structures of sign language. By first predicting gloss and translation sequences and then performing text-based QA, Sign2Text2Answer substantially outperforms the general-purpose VLMs, demonstrating the importance of in-domain CSLR and SLT knowledge. Nevertheless, SLQAM consistently achieves the best performance across gloss recognition, structured QA, translation understanding, and overall metrics. This suggests that directly conditioning visual representations on the question is more effective than relying on explicitly predicted intermediate sequences, which may introduce error propagation in the cascaded pipeline.

Evaluation of SLQA Performance on CSL-Daily-QA. Tables 7 and 8 compare SLQAM with general-purpose VLMs and the cascaded Sign2Text2Answer baseline on CSL-Daily-QA dataset. Consistent with the results on PHOENIX14T-QA, general-purpose VLMs perform poorly, particularly on gloss recognition and structured QA, indicating their limited ability to capture fine-grained sign articulations and temporal relationships. Sign2Text2Answer substantially outperforms VLMs by leveraging predicted gloss and translation sequences, demonstrating the importance of in-domain sign-language representations. Nevertheless, SLQAM achieves the best overall performance and consistently higher results. These results suggest that direct question-conditioned video modeling can better preserve question-relevant visual information and mitigate the error propagation caused by explicitly predicted intermediate representations.

7 Conclusion

In this paper, we introduced a new SL understanding task, *Sign Language Question Answering* (SLQA), which aims to evaluate sign language understanding through natural language question answering rather than fixed prediction objectives. To facilitate this task, we constructed two SLQA benchmarks based on the widely used PHOENIX14T and CSL-Daily datasets by leveraging their existing gloss and translation annotations. The resulting benchmarks cover five complementary question categories, including *Position Reasoning*, *Structural Reasoning*, *Visual Search*, *Gloss Recognition*, and *Translation Understanding*, providing a more comprehensive evaluation of SLU. Furthermore, we proposed a simple yet effective baseline model, SLQAM, equipped with two key components: (1) an in-domain knowledge transfer strategy that progressively transfers knowledge from CSLR and SLT to SLQA, and (2) a QCMTD module that dynamically selects question-relevant temporal features for answer generation. Extensive experiments on both benchmarks demonstrate that the proposed baseline consistently outperforms existing general-purpose video question answering models, highlighting the unique challenges of SLQA and the effectiveness of the proposed framework. We hope that SLQA will serve as a standardized benchmark for evaluating SLU beyond conventional recognition and translation tasks, and inspire future research on more general and reasoning-oriented SL foundation models.

AI USE STATEMENT

Generative AI tools (ChatGPT) were used in the preparation of this manuscript only for language editing, including improving grammatical correctness, readability, and clarity, as well as identifying typographical and writing errors. They were not involved in the formulation of the SLQA task, the design or construction of the proposed benchmarks, the generation or annotation of question–answer pairs, the development of the proposed method, code implementation, experimental design and execution, or result analysis. These components were carried out entirely by the authors.

All AI-assisted edits were manually reviewed by the authors to ensure that they preserve the intended technical meaning and do not introduce new scientific claims or unsupported content. The authors take full responsibility for all content presented in this work.

ETHICS STATEMENT

The SignQA benchmarks introduced in this work are constructed entirely from existing publicly available sign language datasets. We do not collect new videos, recruit additional participants, or acquire any additional personal information. The question–answer pairs are derived from the original annotations through predefined templates and are used solely for research purposes. We follow the usage conditions of the source datasets and do not introduce additional privacy risks beyond those associated with the original datasets.

As SignQA inherits the signer populations, linguistic coverage, and data distributions of its source datasets, it may also inherit their existing biases and limitations. Models evaluated on SignQA should therefore not be assumed to generalize uniformly across different signers, sign languages, or real-world settings. The proposed benchmark is intended for research on sign language understanding, and we encourage appropriate evaluation and validation before applying models trained on these datasets to real-world applications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the SignQA benchmark construction, question–answer generation procedure, proposed baseline, training objectives, and experimental settings in the main paper and Appendix. The SignQA benchmarks will be made publicly available to facilitate future research and reproducibility. All comparison models are implemented and trained based on publicly available codebases, enabling independent verification of the reported results. Additional implementation details, training configurations, and hyperparameter settings are provided in the Appendix. The source code for benchmark construction, model training, and evaluation will be publicly released upon acceptance. Together, these resources provide the necessary information to reproduce the benchmark construction and experimental results reported in this work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pp. 7784–7793, 2018.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*, pp. 5120–5130, 2022a.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022b.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7071–7081, 2024.

Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV*. Springer, 2020.

Edward Fish and Richard Bowden. Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. *arXiv preprint arXiv:2506.00129*, 2025.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Skeleton-aware neural sign language translation. In *MM*, pp. 4353–4361, 2021.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Towards real-time sign language recognition and translation on edge devices. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 4502–4512, 2023.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Hongkai Wen, Lei Xie, and Sanglu Lu. Signgraph: A sign sequence is worth graphs of nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13479, 2024.

Shiwei Gan, Xiao Liu, Yafeng Yin, Zhiwei Jiang, Bowen Guo, Lie Xie, Sanglu Lu, and Hongkai Wen. Signllama: Enhancing gloss-free sign language translation by prioritizing visual features for llms. *Proceedings of the ACM international conference on multimedia*, 2026a.

Shiwei Gan, Xiao Liu, Yafeng Yin, Nan Liu, Kuizhuang Liu, Desibieer Tuerdaken, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Learning effective sign features without text for gloss-free sign language translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9827–9836, 2026b.

Shiwei Gan, Lichen Wang, Xiao Liu, Yafeng Yin, Kuizhuang Liu, Sanglu Lu, and Lei Xie. Sign language question answering: A new task, benchmark, and baseline for sign language understanding. *arXiv preprint arXiv:2607.27826*, 2026c.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Mixsigngraph: A sign sequence is worth mixed graphs of nodes. *Advances in Neural Information Processing Systems*, 38:139371–139395, 2026d.

Liqing Gao, Fan Lyu, Peng Shi, Lei Zhu, Junfu Pu, Liang Wan, and Wei Feng. Overcoming modality bias in question-driven sign language video translation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11724–11738, 2024.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. *Findings of the Association for Computational Linguistics: ACL 2025*, 2025a.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28792–28810, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.1397/.

Dan Guo, Wengang Zhou, Meng Wang, and Houqiang Li. Sign language recognition based on adaptive hmms with data augmentation. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2876–2880. IEEE, 2016.

Jianyuan Guo, Peike Li, and Trevor Cohn. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *arXiv preprint arXiv:2505.15438*, 2025.

Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *ICCV*, pp. 11303–11312, 2021.

Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. Global-local enhancement network for nmfs-aware sign language recognition. *arXiv preprint arXiv:2008.10428*, 2020.

Hezhen Hu, Wengang Zhou, and Houqiang Li. Hand-model-aware sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1558–1566, 2021.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *European conference on computer vision*, pp. 511–527. Springer, 2022.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 854–862, 2023.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018.

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, 2024.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *ICCV*, pp. 20676–20686, 2023.

Peiqi Jiao, Yuecong Min, and Xilin Chen. Visual alignment pre-training for sign language translation. In *European Conference on Computer Vision*, pp. 349–367. Springer, 2024.

Tao Jin and Zhou Zhao. Contrastive disentangled meta-learning for signer-independent sign language translation. In *MM*, pp. 5065–5073, 2021.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3367–3376, 2022.

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6205–6214, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024.

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

Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Yi Yang, et al. Gloss-free end-to-end sign language translation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Tao Liu, Wengang Zhou, and Houqiang Li. Sign language recognition with long short-term memory. In *ICIP*, pp. 2871–2875. IEEE, 2016.

Xiao Liu, Shiwei Gan, Yafeng Yin, Bowen Guo, Zhiwei Jiang, Shunmei Meng, Lei Xie, and Sanglu Lu. Signpr: A progressive vector-quantized diffusion framework for sign language production. pp. 2198–2208, 2026a.

Xiao Liu, Shiwei Gan, Yafeng Yin, Jiaxin Yin, Bowen Guo, Yaqi Sun, Zhiwei Jiang, and Lei Xie. Bridging the gap between semantics and reconstruction: Unifying sign language translation and production. *Proceedings of the 31st ACM international conference on multimedia*, 2026b.

Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. Scope: Sign language contextual processing with embedding from llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5739–5747, 2025.

Zekang Liu, Wei Feng, Fanhua Shang, Lianyu Hu, Jichao Feng, and Liqing Gao. Ssl-ssaw: Self-supervised learning with sigmoid self-attention weighting for question-based sign language translation. *Pattern Recognition*, pp. 114189, 2026c.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, pp. 11542–11551, 2021.

Yuecong Min, Yifan Yang, Peiqi Jiao, Zixi Nan, and Xilin Chen. A closer look at skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4909–4915, 2025.

Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *ECCV*, pp. 172–186. Springer, 2020.

Alptekin Orbay and Lale Akarun. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 222–228. IEEE, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. Spatio-temporal graph convolutional networks for continuous sign language recognition. In *ICASSP*. IEEE, 2022.

Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *MM*, pp. 1497–1505, 2020.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. Towards privacy-aware sign language translation at scale. *arXiv preprint arXiv:2402.09611*, 2024a.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8624–8641, Bangkok, Thailand, 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.467.

Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *TMM*, 2021.

Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. *arXiv preprint arXiv:2407.11144*, 2024.

Zeno Testa, Antonino Furnari, Lorenzo Baraldi, and Natalia Díaz-Rodríguez. Slu-2k: A question-based benchmark for semantic evaluation of sign language translation. *arXiv preprint arXiv:2606.03788*, 2026.

Dominique Uebersax, Juergen Gall, Michael Van den Bergh, and Luc Van Gool. Real-time sign language letter and word recognition from depth data. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pp. 383–390. IEEE, 2011.

Dave Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36: 29029–29047, 2023.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

Zhihao Wang, Shiyu Liu, Zhiwei He, Kangjie Zheng, Liangying Shao, Junfeng Yao, and Jinsong Su. Gloss matters: Unlocking the potential of non-autoregressive sign language translation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4127–4136, 2025.

Chengcheng Wei, Jian Zhao, Wengang Zhou, and Houqiang Li. Semantic boundary detection with reinforcement learning for continuous sign language recognition. *TCSVT*, 31(3):1138–1149, 2020.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: The Twelfth International Conference on Learning Representations*, 2024.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Signrep: Enhancing self-supervised sign representations. *arXiv preprint arXiv:2503.08529*, 2025.

Zhixuan Wu, Bo Cheng, Jiale Han, Jiabao Ma, Shuhao Zhang, Yuli Chen, and Changbo Li. Videoqata: Temporal-aware multi-modal video question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7239–7252, 2025.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, and Yu-Wing Tai. Sf-net: Structured feature network for continuous sign language recognition. *arXiv preprint arXiv:1908.01341*, 2019.

Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simulslt: End-to-end simultaneous sign language translation. In *MM*, pp. 4118–4127, 2021.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Zhihao Zhang, Junfu Pu, Liansheng Zhuang, Wengang Zhou, and Houqiang Li. Continuous sign language recognition via reinforcement learning. In *ICIP*, pp. 285–289. IEEE, 2019.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the aaai conference on artificial intelligence*, volume 38, pp. 19643–19651, 2024.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3597–3605, 2023.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 6439–6455, 2022.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *ICCV*, pp. 20871–20881, 2023.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, pp. 1316–1325, 2021.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. Scaling up multimodal pre-training for sign language understanding. *TPAMI*, 2025.

Qidan Zhu, Jing Li, Fei Yuan, and Quan Gan. Multiscale temporal network for continuous sign language recognition. *Journal of Electronic Imaging*, 33(2):023059–023059, 2024.

Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, pp. 5131–5140, 2022.

Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14890–14900, 2023.

Table 9: SLQA datasets statistics. One question per module per video; splits follow the official corpus partitions (train/dev/test: 35,480/2,595/3,210 for PHOENIX-2014T and 92,000/5,385/5,880 for CSL-Daily). †M3 includes one CSL-only non-manual (interrogative) template (3,448 QA); PHOENIX-2014T is declarative-only.

Question Category	#Templates	PHOENIX-2014T			CSL-Daily		
		Train	Dev	Test	Train	Dev	Test
M1 Position Reasoning	9	7,096	519	642	18,400	1,077	1,176
M2 Structural Reasoning	13	7,096	519	642	18,400	1,077	1,176
M3 Visual Search†	5(+1)	7,096	519	642	18,400	1,077	1,176
M4 Gloss Recognition	1	7,096	519	642	18,400	1,077	1,176
M5 Translation Understanding	1	7,096	519	642	18,400	1,077	1,176
Total QA Pairs	**30**	35,480	2,595	3,210	92,000	5,385	5,880

Table 10: Representative question–answer pairs for each question category in SLQA. [de] denotes PHOENIX-2014T, and [zh] denotes CSL-Daily. Long answers are truncated for brevity. M3 includes both manual gloss localization ($M3_{man.}$) and non-manual interrogativity detection ($M3_{n.m.}$).

Module	Representative Question–Answer Pair *(English translation)*
M1	**[de]** Welches Gloss steht an Position 2? → "ZUSCHAUER". *(Which gloss appears at position 2? →* VIEWER.*)*
M2	**[zh]** 第一个动作之后的那个是什么？ → 第2个 gloss 为"时间"。 *(Which gloss comes immediately after the first sign? → the second gloss is* TIME.*)*
$M3_{man.}$	**[de]** Erstes Vorkommen von "KUESTE"? → Position 6. *(Where does* COAST *first appear? → position 6.)*
$M3_{n.m.}$	**[zh]** 结合面部表情，这是疑问句吗？ → 不是疑问句。 *(Based on the non-manual markers, is this an interrogative sentence? → No.)*
M4	**[de]** Gloss-Sequenz des Videos? → "TEIL BISSCHEN FROST NACHT . . .". *(What is the gloss sequence of the video? →* PART BIT FROST NIGHT . . .*)*
M5	**[zh]** 完整句子是什么？ → "这些经典的文学作品他全都读过"。 *(What is the complete spoken-language sentence? → "He has read all these classic works.")*

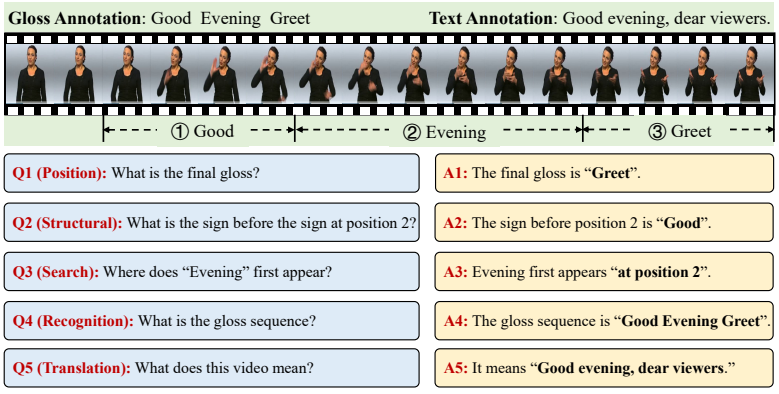


Figure 3: Illustration of each question category in SLQA.

Table 11: Effect of the proposed in-domain knowledge transfer strategy on CSL-daily-QA.

Training Strategy	Gloss Recognition		Structured QA		Non Manual		Translation Understanding (Dev)						Translation Understanding (Test)					
	WER↓		ACC↑		F1↑		$BLEURT_{SLT}$	RL	B1	B2	B3	B4	$BLEURT_{SLT}$	RL	B1	B2	B3	B4
	Dev	Test	Dev	Test	Dev	Test												
+ Stage 1 (CSLR)	34.17	33.67	64.23	64.36	90.43	92.15	57.56	53.27	52.75	39.33	29.86	23.21	58.32	53.67	52.74	39.54	30.27	23.67
+ Stage 2 (SLT)	34.19	34.01	62.56	62.71	91.49	93.19	58.00	52.99	52.99	39.71	30.43	23.95	57.87	53.16	52.56	39.43	30.30	23.84

Table 12: Overall QA performance under different training strategies on CSL-daily-QA.

Training Strategy	Overall Dev							Overall Test						
	RL	B1	B2	B3	B4	BLEURT	CIDEr	RL	B1	B2	B3	B4	BLEURT	CIDEr
+ Stage 1 (CSLR)	73.07	74.76	68.35	62.98	58.81	60.27	4.150	73.46	75.20	68.79	63.44	59.02	60.51	4.212
+ Stage 2 (SLT)	73.39	75.31	68.93	63.60	59.16	60.41	4.179	73.52	75.32	68.92	63.60	59.18	60.23	4.193

A Experimental Details

Datasets. We evaluate the proposed method on the SignQA benchmark, which is constructed from two widely used sign language translation datasets: **PHOENIX-2014T** (Camgoz et al., 2018) and **CSL-Daily** (Zhou et al., 2021). PHOENIX-2014T is a German Sign Language (DGS) dataset containing 7,096 training, 519 validation, and 642 test videos collected from nine signers. CSL-Daily is a Chinese Sign Language dataset consisting of 18,400 training, 1,077 validation, and 1,176 test videos from ten signers.

Following the dataset construction procedure described in Section **Dataset Construction**, we generate one question for each of the five question categories from every video, resulting in five question–answer pairs per sample. We strictly follow the official train/validation/test splits of the original datasets, ensuring that no additional data leakage is introduced. In total, SignQA contains 41,285 QA pairs for PHOENIX-2014T (35,480/2,595/3,210 for train/validation/test) and 103,265 QA pairs for CSL-Daily (92,000/5,385/5,880 for train/validation/test), yielding a total of 144,550 question–answer pairs. The benchmark covers five complementary question categories, namely *Position Reasoning*, *Structural Reasoning*, *Visual Search*, *Gloss Recognition*, and *Translation Understanding*. Detailed statistics are reported in Table 9.

Data Preprocessing. We adopt following augmentations, including resizing to 256×256 pixels, random cropping to 224×224 pixels, random horizontal flipping (probability 0.5), and random temporal scaling within ±20%. During inference, frames are resized to 256×256 pixels and center-cropped to 224×224 pixels.

B Examples of the Constructed SLQA Datasets

Table 10 and Figure 3 present representative question–answer examples from the constructed SignQA datasets. Unlike previous question-guided SLT tasks, where questions primarily serve as auxiliary conditions for generating different translations, our questions directly probe the content of the sign language video. In particular, SignQA covers five complementary types of questions: *Position Reasoning*, *Structural Reasoning*, *Visual Search*, *Gloss Recognition*, and *Translation Understanding*.

As illustrated in the table, these question types require models to operate at different levels of sign language understanding. *Position Reasoning* and *Structural Reasoning* evaluate whether a model can identify and reason over the ordered relationships among signs. *Visual Search* requires the model to locate specific signs in the video and, where applicable, interpret non-manual markers such as facial expressions. *Gloss Recognition* evaluates fine-grained sign-level perception, while *Translation Understanding* assesses the ability to infer the overall spoken-language meaning conveyed by the video. Together, these tasks cover both fine-grained visual and temporal understanding and high-level semantic interpretation, providing a more comprehensive evaluation of sign language understanding than translation-only objectives.

Table 13: Qualitative comparison of **Position Reasoning** results on PHOENIX14T-QA and CSL-Daily-QA.

PHOENIX14T-QA	
Question	Bestimmen Sie das finale Gloss-Token.
Ground Truth	Das Schluss-Gloss lautet “REGEN”.
VideoLLaMA3-2B	Das Schluss-Gloss lautet “GRAD”.
Qwen3-VL-2B-Instruct	Das Schluss-Gloss lautet “REGEN”.
InternVL3-2B	Das Schluss-Gloss lautet “REGEN”.
SLQAM (ours)	Das Schluss-Gloss lautet “REGEN”.
CSL-Daily-QA	
Question	请确定序列末尾的终止手语词。
Ground Truth	序列结尾的gloss是“运动”。
VideoLLaMA3-2B	最后的gloss为“不”。
Qwen3-VL-2B-Instruct	最后的gloss为“可以”。
InternVL3-2B	序列结尾的gloss是“跑步”。
SLQAM (ours)	序列结尾的gloss是“运动”。

C Ablation Study on CSL-daily-QA

We further evaluate the proposed in-domain knowledge transfer strategy on CSL-Daily-QA. The category-level and overall results are reported in Tables 11 and 12, respectively. Compared with Stage 1 (CSLR pre-training), introducing Stage 2 (SLT pre-training) yields clear improvements on the *Non-Manual* question type, increasing F1 from 90.43 to 91.49 on the development set and from 92.15 to 93.19 on the test set. Stage 2 also improves most translation metrics on the development set, including $\text{BLEURT}_{\text{SLT}}$, BLEU-2, BLEU-3, and BLEU-4, indicating that SLT pre-training provides additional sentence-level semantic knowledge and benefits language generation.

As shown in Table 12, Stage 2 further improves the overall QA performance on most lexical metrics. These improvements demonstrate that SLT pre-training complements the sign-specific visual and temporal representations learned during CSLR pre-training by strengthening cross-modal semantic alignment and answer-generation capability. Nevertheless, the gains are not uniform across all categories. Gloss recognition performance remains nearly unchanged, while structured QA accuracy decreases slightly after Stage 2. Moreover, BLEURT and CIDEr show small fluctuations on the test set. This suggests that CSLR pre-training already provides a strong foundation for fine-grained recognition and structural reasoning, whereas SLT pre-training mainly benefits sentence-level semantics and generation-oriented metrics. Overall, the results on CSL-Daily-QA confirm the complementary roles of the two training stages: CSLR pre-training establishes sign-specific visual representations, while SLT pre-training further enhances semantic alignment and language generation.

D Qualitative Analysis

We present representative qualitative examples from the two constructed SignQA datasets in Tables 13–17, covering *Position Reasoning*, *Structural Reasoning*, *Visual Search*, *Gloss Recognition*, and *Translation Understanding*. These examples provide a qualitative comparison between general-purpose video-language models and our sign-language-specific baseline.

For the reasoning-oriented tasks, general-purpose VLMs often produce plausible but incorrect answers despite the relatively simple question structures. For example, in the *Structural Reasoning* example, Qwen3-VL and InternVL3 incorrectly predict a gloss before “WIND”, although it is the first gloss in the sequence. Similarly, in the *Visual Search* example, Qwen3-VL predicts an incorrect occurrence position for “SONNTAG”. These errors suggest that general-purpose VLMs may recognize relevant visual or semantic content but struggle to accurately establish the fine-grained temporal and sequential relationships required by SLQA. In contrast, our model correctly identifies the queried positions and local relationships.

Table 14: Qualitative comparison of **Structural Reasoning** results on two benchmarks.

PHOENIX14T	
Question	Welches Gloss erscheint unmittelbar vor “WIND”?
Ground Truth	“WIND” ist das erste Gloss; davor steht nichts.
VideoLLaMA3-2B	Vor “WIND” steht nichts mehr, es ist das erste Gloss.
Qwen3-VL-2B-Instruct	Vor “WIND” kommt “SCHWACH”.
InternVL3-2B	Vor “WIND” kommt “MAESSIG”.
SLQAM (ours)	“WIND” ist das erste Gloss; davor steht nichts.
CSL-Daily-QA	
Question	哪个gloss出现在“好处”之前?
Ground Truth	“好处”的前一个gloss是“好”。
VideoLLaMA3-2B	“好处”前面紧挨着“这”。
Qwen3-VL-2B-Instruct	在“好处”之前的是“有”。
InternVL3-2B	“好处”前面紧挨着“有”。
SLQAM (ours)	“好处”的前一个gloss是“好”。

Table 15: Qualitative comparison of **Visual Search** results on two benchmarks.

PHOENIX14T	
Question	Wann passiert “SONNTAG” zum letzten Mal?
Ground Truth	Das letzte Auftreten von “SONNTAG” ist an Position 1.
VideoLLaMA3-2B	Das letzte Auftreten von “SONNTAG” ist an Position 1.
Qwen3-VL-2B-Instruct	Das letzte Auftreten von “SONNTAG” ist an Position 2.
InternVL3-2B	Die letzte Stelle für “SONNTAG” ist Position 1.
SLQAM (ours)	Das letzte Auftreten von “SONNTAG” ist an Position 1.
CSL-Daily-QA	
Question	“旁边”最后一次出现是在第几步?
Ground Truth	“旁边”最近一次出现的位置是第2位。
VideoLLaMA3-2B	“旁边”最后一次出现在第3位。
Qwen3-VL-2B-Instruct	“旁边”最近一次出现的位置是第4位。
InternVL3-2B	“旁边”最后一次出现于位置3。
SLQAM (ours)	“旁边”最近一次出现的位置是第2位。

The *Gloss Recognition* example further highlights the importance of fine-grained sign-level visual understanding. While the general-purpose VLMs produce gloss sequences with similar high-level content, they make errors in fine-grained details such as dates and numbers. These seemingly small recognition errors can substantially affect the correctness of the entire gloss sequence. Our model, benefiting from in-domain CSLR knowledge, correctly recovers the complete gloss sequence in this example.

The difference is also evident in *Translation Understanding*. The general-purpose VLMs generate fluent German sentences, but their answers describe weather events that differ from those conveyed by the sign video. This indicates that linguistic fluency alone does not guarantee accurate sign-to-language semantic understanding. In contrast, our model preserves the key semantic content of the video, including the temporal and spatial information as well as the weather condition. Overall, these examples illustrate that the proposed SLQA model is better able to connect fine-grained visual evidence with sequential reasoning and language generation, rather than merely producing plausible responses based on general video-language priors.

E Impact Statement

This work introduces a new task, Sign Language Question Answering (SLQA), together with two benchmark datasets and a baseline model. We believe this work has the potential to advance sign language understanding by enabling more flexible and interactive access to sign language videos. Compared with conventional sign language recognition and translation systems that produce fixed

Table 16: Qualitative comparison of **Gloss Recognition** results on two benchmarks.

PHOENIX14T-QA	
Question	Schreiben Sie die Gebärdenwörter, die Sie sehen, der Reihe nach auf.
Ground Truth	Die Gloss-Sequenz des Videos ist “MORGEN WETTER WIE-AUSSEHEN DIENSTAG FUENFTE JANUAR”.
VideoLLaMA3-2B	Die Gloss-Sequenz des Videos ist “MORGEN WETTER WIE-AUSSEHEN DIENSTAG DREIZEHN FEBRUAR”.
Qwen3-VL-2B-Instruct	Die Gloss-Sequenz des Videos ist “MORGEN WETTER WIE-AUSSEHEN DIENSTAG VIERZEHN MAI”.
InternVL3-2B	Die Gloss-Sequenz des Videos ist “MORGEN WETTER WIE-AUSSEHEN DIENSTAG NEUNTE FEBRUAR”.
SLQAM (ours)	Die Gloss-Sequenz des Videos ist “MORGEN WETTER WIE-AUSSEHEN DIENSTAG FUENFTE JANUAR”.
CSL-Daily-QA	
Question	把看到的动作代号按顺序写下来。
Ground Truth	这个手语视频的gloss序列是“我 丈夫 在 美国 中国 菜 喜欢”。
VideoLLaMA3-2B	这个手语视频的gloss序列是“我 为 明天 跳舞 准备 音乐”。
Qwen3-VL-2B-Instruct	这个手语视频的gloss序列是“我 买 好了 面包 牛奶”。
InternVL3-2B	这个手语视频的gloss序列是“我 丈夫 每天 火锅 喜欢 吃”。
SLQAM (ours)	这个手语视频的gloss序列是“我 丈夫 在 美国 中国 菜 喜欢”。

Table 17: Qualitative comparison of **Translation Understanding** results on two benchmarks.

PHOENIX14T-QA	
Question	Was ist der vollständige Satz, der vom Gebärdenden ausgedrückt wird?
Ground Truth	Diese Gebärdsequenz bedeutet “am donnerstag regnet es vor allem im süden und osten mancherorts”.
VideoLLaMA3-2B	Diese Gebärdsequenz bedeutet “am tag regnet es im süden und osten teilweise kräftig sonst ist es meist freundlich”.
Qwen3-VL-2B-Instruct	Diese Gebärdsequenz bedeutet “am samstag im norden und westen viele wolken hier und da regnet oder schauert es”.
InternVL3-2B	Diese Gebärdsequenz bedeutet “am sonntag ist es im süden freundlich”.
SLQAM (ours)	Diese Gebärdsequenz bedeutet “am donnerstag regnet es vor allem im süden und südosten teilweise kräftig”.
CSL-Daily-QA	
Question	手语者表达的完整句子是什么?
Ground Truth	这段手语的意思是“你是我的客户,我应该请你吃饭.”。
VideoLLaMA3-2B	这段手语的意思是“他把手机里的自拍照都删掉了.”。
Qwen3-VL-2B-Instruct	这段手语的意思是“他把多余的钱都存到了银行.”。
InternVL3-2B	这段手语的意思是“你让我收拾房间,我怎么有种吃不消的感觉啊!”。
SLQAM (ours)	这段手语的意思是“你是我的客户,我应该请你吃饭.”。

outputs, SLQA allows users to query different aspects of sign language videos through natural language, which may benefit applications such as sign language education, intelligent retrieval, and assistive communication technologies. As with other data-driven sign language systems, our approach inherits the limitations and potential biases of the underlying datasets, including demographic imbalance, limited signer diversity, and constrained recording conditions. Consequently, the model may exhibit degraded performance for unseen signing styles, rare linguistic phenomena, or challenging visual environments. Therefore, the proposed system should not be deployed without appropriate human oversight in safety-critical or high-stakes applications where recognition errors could have significant consequences. Overall, we believe that improving the flexibility and accessibility of sign language understanding can provide meaningful societal benefits, provided that the technology is developed and deployed responsibly.

F LIMITATIONS AND DISCUSSION

We outline several limitations of the proposed method and discuss potential directions for future work.

Limited Dataset Coverage. The proposed SignQA datasets are constructed from PHOENIX14T and CSL-Daily, both of which focus on relatively constrained domains and controlled recording conditions. Therefore, they cannot fully capture the linguistic diversity, signer variability, and environmental complexity of sign language videos in real-world scenarios. As a result, models trained on these datasets may exhibit limited generalization to unseen domains, signing styles, and recording conditions. In future work, we plan to expand SignQA by incorporating larger and more diverse sign language datasets, with the goal of improving model robustness and enabling more reliable interaction with sign language videos in real-world applications.

Three-Stage Training. Our baseline model relies on a three-stage training pipeline to transfer in-domain knowledge from CSLR and SLT to SLQA. Although this strategy substantially improves performance, it inevitably increases the overall training time, computational cost, and implementation complexity. Developing a more efficient unified training strategy that can jointly learn sign recognition, translation, and question-answering capabilities remains an important direction for future research.

Scope of the SLQA Tasks. SignQA is constructed from existing gloss and translation annotations, and therefore its questions primarily evaluate information already represented in these annotation layers. M4 and M5 correspond closely to conventional CSLR and SLT, while M1–M3 apply controlled retrieval and relational operations to gloss-level content. Consequently, strong performance on SignQA should not be interpreted as evidence of complete or human-level sign-language understanding. In particular, the benchmark does not fully capture simultaneity, spatial grammar, classifier constructions, discourse structure, or other linguistic phenomena that may not be represented in linear gloss annotations. We view SignQA as a controllable diagnostic complement to existing SLU benchmarks.

We hope this benchmark will support more transparent analysis of how sign-language models retrieve and use different types of signed information under natural-language queries. We therefore view the current version of SignQA as an initial, controllable diagnostic benchmark rather than a comprehensive evaluation of sign-language understanding. Future versions will incorporate additional human-annotated question types that require reasoning over linguistic phenomena beyond those represented in the existing gloss and translation annotations.