

# CNSL-bench: Benchmarking the Sign Language Understanding Capabilities of MLLMs on Chinese National Sign Language

Rui Zhao<sup>1,2,3</sup>, Xuewen Zhong<sup>1,2,3</sup>, Xiaoyun Zheng<sup>1,2,3</sup>,  
Jinsong Su<sup>1,2</sup> and Yidong Chen<sup>1,2,3\*</sup>

<sup>1</sup>School of Informatics, Xiamen University, China

<sup>2</sup>Key Lab of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian-Taiwan (XMU), Ministry of Culture and Tourism, China

<sup>3</sup>National Language Resources Monitoring and Research Center for Education and Teaching Media, Xiamen University, China  
zhsqzr@stu.xmu.edu.cn ydchen@xmu.edu.cn

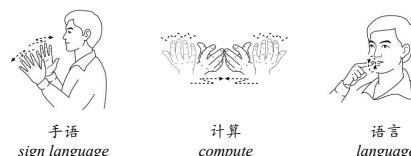
## Abstract

Sign language research has achieved significant progress due to the advances in large language models (LLMs). However, the intrinsic ability of LLMs to understand sign language, especially in multimodal contexts, remains underexplored. To address this limitation, we introduce **CNSL-bench**, the first comprehensive Chinese National Sign Language benchmark designed for evaluating multimodal large language models (MLLMs) in sign language understanding. The proposed CNSL-bench is characterized by: 1) Authoritative grounding, as it is anchored to the officially standardized *National Common Sign Language Dictionary*, mitigating ambiguity from regional or non-canonical variants and ensuring consistent semantic definitions; 2) Multimodal coverage, providing aligned textual descriptions, illustrative images, and sign language videos; and 3) Articulatory diversity, supporting fine-grained analysis across key manual articulatory forms, including air-writing, finger-spelling, and the Chinese manual-alphabet. Using CNSL-bench, we extensively evaluate 21 open-source and proprietary up-to-date MLLMs. Our results reveal that, despite recent advances in multimodal modeling, current MLLMs remain substantially inferior to human performance, exhibiting systematic disparities across input modalities and manual articulatory forms. Additional diagnostic analyses suggest that several performance limitations persist beyond improvements in reasoning and that instruction-following robustness varies substantially across models.

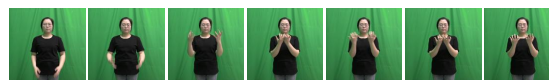
## 1 Introduction

Sign language plays a central role in communication for many people with hearing impairment and has consequently attracted sustained attention from the research community over the past decades.

\*Corresponding Author.



双手直立，掌心左右相对，前后交替移动几下，表示打手语。  
Stand with both hands upright, palms facing each other on the left and right, and move them forward and backward alternately a few times to indicate signing.



双手五指微曲，掌心向上，边交替点动边互碰两下。  
Bend the fingers of both hands slightly, with the palms facing up. Alternate tapping them while touching each other twice.



一手食指横伸，在嘴前前后转动两下。  
Extend the index finger of one hand horizontally and rotate it twice in front of the mouth.

Figure 1: Examples from CNSL-bench showing the aligned textual description, illustrative image, and corresponding sign language video for each sign entry.

More recently, advances in large language models (LLMs) have further stimulated progress in automatic sign language understanding, primarily within specific tasks, e.g., sign language translation, where LLMs are incorporated as semantic augmentation modules or enhanced text decoders for improved performance (Wong et al., 2023; Gong et al., 2024; Chen et al., 2024b; Guo et al., 2025; Kim et al., 2025; Liu et al., 2025; Jang et al., 2025; Asasi et al., 2025; Rao et al., 2025; Hwang et al., 2025).

Despite encouraging task-level improvements, existing approaches predominantly embed LLMs into downstream pipelines or datasets, leaving the intrinsic ability of these models to understand sign language largely unexamined. This limitation becomes even more pronounced in the context of multimodal large language models (MLLMs), which have demonstrated strong visual-language capa-

bilities over images and videos (Liu et al., 2023, 2024; Zhang et al., 2024; Shen et al., 2025). Crucially, sign language possesses the full range of fundamental linguistic properties and is inherently multimodal, with meaning expressed through the coordinated use of linguistically grounded manual articulatory cues such as air writing, finger-spelling, and the manual-alphabet (Shi et al., 2021; Yin et al., 2021; Desai et al., 2024; Atwell et al., 2024). While this rich expressiveness poses fundamental challenges for automatic sign language understanding, it also raises an open question: to what extent can current MLLMs genuinely comprehend sign language by capturing linguistically grounded structure and semantic meanings rather than relying solely on visual correlations?

To answer this question, we introduce **CNSL-bench**, the first comprehensive Chinese National Sign Language benchmark designed for evaluating MLLMs in sign language understanding (§2.1). CNSL-bench is constructed upon the *National Common Sign Language Dictionary*, an officially standardized lexical resource for Chinese national sign language. This authoritative grounding provides a canonical semantic reference, reducing ambiguity from regional or variants and enabling consistent, controlled evaluation of sign language understanding (Ministry of Education of the People’s Republic of China et al., 2018b; China Disabled Persons’ Federation et al., 2019). To enable multimodal evaluation, we further align these dictionary entries with a large-scale video dataset covering isolated Chinese national sign language (Jin et al., 2025), resulting in a unified benchmark that provides broad multimodal coverage, where each sign entry is represented by aligned *textual description*, *illustrative image*, and *sign language video*, as exemplified in Figure 1. As a consequence, CNSL-bench comprises 20,121 questions spanning text, image, and video. Furthermore, the benchmark explicitly supports manual articulatory diversity, covering three representative categories of signs, including air-writing, finger-spelling, and manual-alphabet, as shown in Figure 2. These categories are treated as dedicated evaluation subsets to enable fine-grained analysis. Supported by high-quality resources, carefully designed evaluation protocols, and human assessment involving the Deaf<sup>1</sup> community, CNSL-bench serves as a reliable and compre-

<sup>1</sup>We follow the recognized convention of using the upper-cased word Deaf to refer to the community of sign language users (Woodward, 1972)

hensive benchmark for systematically diagnosing the intrinsic sign language understanding capabilities of modern MLLMs.

With CNSL-bench, we benchmark a wide range of open- and closed-source up-to-date MLLMs, offering a systematic empirical assessment of their sign language understanding capabilities (§ 3.2). Our main results reveal several patterns: **(a)** Despite recent advances in multimodal modeling, current MLLMs remain substantially inferior to human performance in sign language understanding. **(b)** A pronounced modality-dependent performance imbalance is observed, with models exhibiting substantially weaker performance on visual inputs compared to text. **(c)** MLLMs demonstrate uneven comprehension across different manual articulatory forms, achieving relatively stronger performance on finger-spelling than on air-writing and the specific manual-alphabet. **(d)** The performance gap between open-source and proprietary MLLMs is rapidly narrowing, with several open-source models achieving performance comparable to lightweight commercial systems. Beyond the main results, additional diagnostic analyses are conducted to further investigate the intrinsic sign language understanding capabilities of current MLLMs (§ 3.3). Test-time scaling via explicit reasoning mechanisms is examined as a diagnostic tool, revealing heterogeneous gains across models and modalities while failing to fundamentally resolve the observed limitations. Prompt token effects and instruction-following robustness are further considered as complementary diagnostic factors, providing additional insight into the behavioral characteristics of current MLLMs.

In summary, the contributions of this work are as follows:

- (1) We introduce **CNSL-bench**, a multimodal, sign language-centric benchmark for evaluating sign language understanding in MLLMs.
- (2) We present a comprehensive evaluation of up-to-date MLLMs on the proposed CNSL-bench.
- (3) Through extensive experiments and analyses, we identify persistent challenges in current MLLMs’ sign language understanding capabilities.

We hope that CNSL-bench will serve as a diagnostic foundation and a reference resource for future research toward more robust, reliable, and human-aligned MLLMs in sign language understanding. Data and code are available at <https://github.com/rzhao-zhsq/CNSL-bench>.

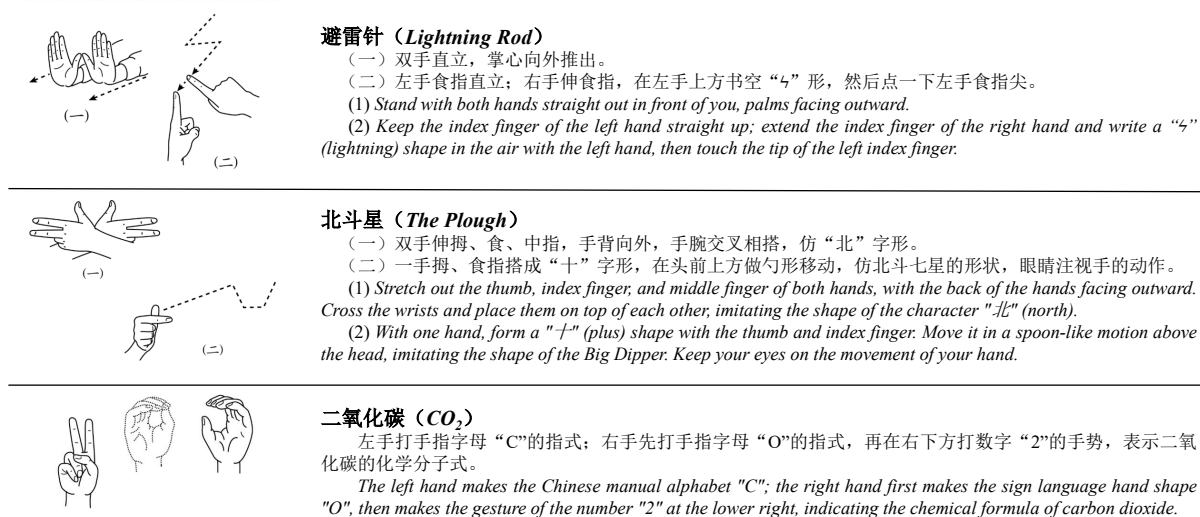


Figure 2: Examples illustrating the three categories of sign articulation: air-writing (top), finger-spelling (middle), and manual-alphabet (bottom).

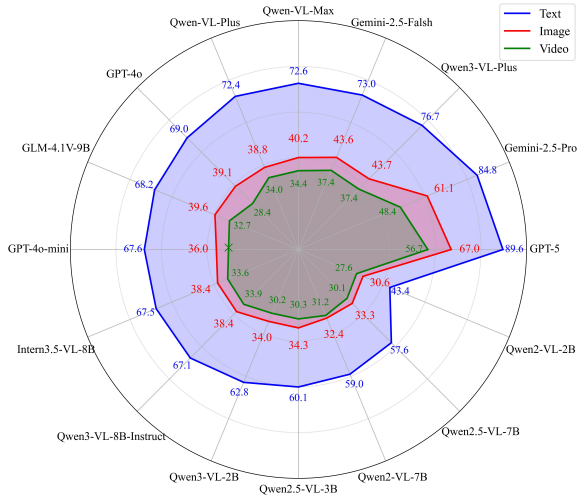
## 2 CNSL-bench

Sign language understanding requires unified modeling of visual, temporal, and fine-grained linguistic information across heterogeneous modalities. To enable a systematic and controlled evaluation of such capabilities in MLLMs, we construct CNSL-bench, a Chinese National Sign Language benchmark grounded in standardized lexical resources and aligned multimodal representations. CNSL-bench is designed to isolate intrinsic sign language understanding from downstream task-specific factors, providing a consistent evaluation framework across text, image, and video inputs.

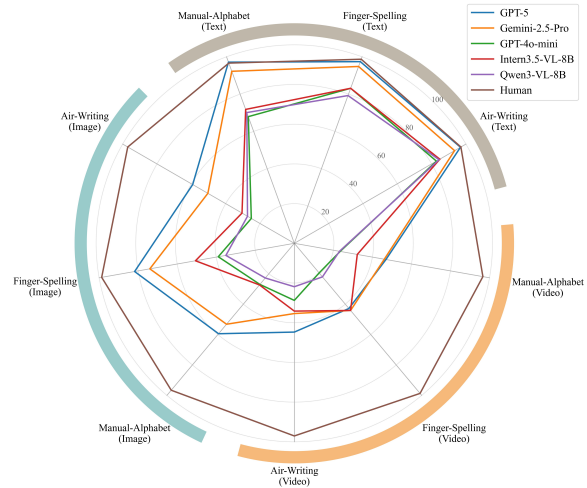
### 2.1 Benchmark Construction

**Benchmark Principle.** CNSL-bench is constructed following three core principles: 1) it is grounded in a standardized lexical foundation. To avoid ambiguity introduced by regional or non-canonical sign variants, the benchmark is anchored to the officially standardized National Common Sign Language Dictionary, ensuring consistent semantic definitions across all samples; 2) it emphasizes aligned multimodal coverage. Each sign entry is systematically represented in text, image, and video formats, enabling controlled evaluation of cross-modal sign language understanding; and 3) the benchmark explicitly incorporates diverse manual articulatory forms, including air-writing, finger-spelling, and specific manual-alphabet signs, supporting fine-grained analysis of linguistically distinct forms in sign language.

**Data Collection.** CNSL-bench is constructed by grounding all entries in officially standardized Chinese national sign language resources and aligning them with multimodal representations. The core lexical inventory is derived from the *National Common Sign Language Dictionary* (China Disabled Persons’ Federation et al., 2019), which is built upon the *Lexicon of Common Expressions in Chinese National Sign Language* jointly issued by the Ministry of Education, the State Language Commission, and the China Disabled Persons’ Federation (Ministry of Education of the People’s Republic of China et al., 2018b). Together, these standards define normative sign realizations that are widely used and stable in education and daily communication. This grounding ensures lexical consistency and reduces regional or informal variation. To ensure that each benchmark item maps to a unique sign entry, we perform systematic sign-level preprocessing to handle dictionary cases where (i) distinct entries share identical hand motions, (ii) identical entries are associated with different meanings or articulations (e.g., “seat belt” for car and airplane), or (iii) the same meaning is realized by distinct hand motions. For each entry, we retain the accompanying textual description and illustrative image, which explain the sign entry for educational and communication purposes. Furthermore, each unique sign entry is aligned with an additional sign language video to reflect real-world usage. The video samples are sourced from a large-scale Chinese national sign language dataset (Jin et al., 2025), yielding a unified multimodal representation



(a) Overall performance on CNSL-bench.



(b) Performance of subsets on CNSL-bench.

Figure 3: A concise summary of the MLLMs’ performance on CNSL-bench. (a) provides a comparison detailing the overall performance of the 16 selected MLLMs, including both open-source and closed-source models. (b) illustrates a radar chart that outlines the performance of humans and MLLMs on three subsets (air-writing, finger-spelling, and manual-alphabet) within CNSL-bench.

for each sign entry.

In addition, CNSL-bench explicitly incorporates specific manual articulatory forms, including air-writing, finger-spelling, and the Chinese manual-alphabet. As exemplified in Figure 2, air-writing refers to tracing graphic forms in the air, and the traced content may correspond to strokes, symbol-like shapes, or a partial character. In our taxonomy, finger-spelling refers to using one or both hands to depict or indicate Chinese character-form structure, prioritizing graphic cues such as outlines, components, and structural patterns over strictly sequential, letter-by-letter spelling. The manual-alphabet follows the *Chinese Manual Alphabet* (Ministry of Education of the People’s Republic of China et al., 2018a), which maps conventionalized finger configurations to individual Chinese Pinyin letters. These letters can be combined to spell Mandarin, forming lexical signs, and functioning as morphemic components within signs under the *Scheme of the Chinese Phonetic Alphabet* (Committee for Language Reform of China, 1957).

The alignment between text, image, and video is detailed in Appendix A.1, and the complete manual alphabets in the *Chinese Manual Alphabet* are listed in Figure 8 (Appendix A.2).

**Task Definition.** For each sign entry, CNSL-bench provides an aligned textual description, an illustrative image, and a sign language video. We formulate the task as a four-way multiple-choice

evaluation, where each instance is instantiated with a single input modality. This closed-form design enables controlled and scalable evaluation, as current state-of-the-art MLLMs remain highly unreliable in open-ended sign language understanding. We further examine different option construction strategies and observe that semantics-based distractors lead to slightly weaker model performance, while yielding conclusions consistent with random sampling. To simplify the benchmark design and facilitate easy reproduction, we therefore adopt random option sampling for robustness and consistency. Detailed task specifications and comparative analyses of option construction strategies, including an open-ended case and distractor design, are provided in Appendix A.3.

## 2.2 Benchmark Statistics

CNSL-bench is grounded in the *National Common Sign Language Dictionary*, which contains 8,214 sign glosses. However, it includes cases where different glosses share identical hand motions, identical glosses correspond to different meanings and articulations, or the same meaning is realized through multiple distinct hand motions. After processing at the sign-entry level, CNSL-bench retains 6,707 unique sign entries, yielding 20,121 evaluation instances across three modalities (text, image, and video). Among the 6,707 sign entries, we manually identify 407 entries containing air-writing, 77 containing finger-spelling, and 592 involving specific

Model	Text				Image				Video <sup>1/2</sup> fps				Video <sup>10</sup> fps			
	AW	FS	MA	All	AW	FS	MA	All	AW	FS	MA	All	AW	FS	MA	All
<i>Open&amp;Close -source Image MLLMs</i>																
LLaVA-NeXT-7B	0.74	2.60	0.68	1.68	17.20	14.29	17.57	17.52	15.72	12.99	17.74	18.67	-	-	-	-
Qwen-VL-Plus	87.71	84.42	77.97	72.41	25.37	28.57	25.89	38.83	17.69	27.27	21.83	32.06	21.38	22.08	23.99	33.95
Qwen-VL-Max	85.50	83.12	77.80	72.62	24.88	33.77	24.87	40.18	20.64	23.38	22.00	31.01	24.08	31.17	23.99	34.43
<i>Open-Source MLLMs</i>																
Qwen2-VL-2B	47.42	51.95	41.55	43.36	19.41	19.48	23.65	30.62	<u>20.39</u>	<b>27.27</b>	<b>22.80</b>	27.23	19.16	18.18	<b>23.65</b>	27.58
Qwen2.5-VL-3B	69.04	72.73	55.41	60.07	<u>25.80</u>	<u>27.27</u>	<u>23.82</u>	<u>34.26</u>	<b>20.64</b>	<u>24.68</u>	17.91	<b>28.36</b>	<b>23.59</b>	<b>28.57</b>	20.61	<u>30.34</u>
Intern-VL-3.5-2B	68.06	<u>79.22</u>	<b>60.64</b>	58.68	19.66	28.57	<b>24.49</b>	<u>33.37</u>	2.95	6.49	4.39	4.29	4.42	6.49	4.39	4.50
Qwen3-VL-2B-Instruct	<b>73.71</b>	<b>83.12</b>	58.61	<b>62.83</b>	21.62	38.96	21.62	34.01	19.66	18.18	<u>21.79</u>	<u>27.78</u>	21.38	19.48	<u>23.14</u>	30.18
Qwen3-VL-2B	66.58	74.03	52.70	57.97	23.59	27.27	20.78	32.80	<u>20.39</u>	16.88	17.74	26.38	<u>22.36</u>	<u>23.38</u>	18.07	28.95
Qwen3-VL-2B $\triangleleft$	<u>72.97</u>	<u>75.32</u>	<u>56.25</u>	<u>61.29</u>	<b>26.54</b>	<b>41.56</b>	21.28	<b>35.37</b>	16.71	15.58	19.76	27.76	<u>22.36</u>	<u>27.27</u>	<u>20.95</u>	<b>30.68</b>
LLaVA-NeXT-Video-7B	1.72	2.60	0.84	1.34	8.85	10.39	12.16	12.94	14.99	18.18	15.20	15.43	13.76	14.29	15.71	15.91
Qwen2-VL-7B	68.30	75.32	56.93	58.98	26.78	31.17	25.17	32.44	<u>22.60</u>	<u>29.87</u>	21.45	29.24	23.10	<u>27.27</u>	23.14	31.19
Qwen2.5-VL-7B	65.60	70.13	55.41	57.61	26.29	27.27	24.83	33.32	21.38	24.68	20.95	29.10	22.11	23.38	21.79	30.15
GLM-4.1V-9B $\triangleleft$	80.34	<b>84.42</b>	69.59	<u>68.24</u>	<u>28.50</u>	<b>55.84</b>	<b>28.38</b>	<b>39.62</b>	20.39	23.38	19.76	28.03	21.87	24.68	21.62	29.75
Intern-VL-3.5-8B	84.77	<u>83.12</u>	<u>71.79</u>	<u>67.53</u>	<b>30.47</b>	<u>50.65</u>	<u>27.36</u>	38.36	<b>34.15</b>	<b>44.16</b>	<b>32.26</b>	<b>32.26</b>	<b>31.70</b>	<b>36.36</b>	<b>29.39</b>	<u>33.59</u>
Qwen3-VL-8B-Instruct	84.28	79.22	70.10	67.06	27.27	35.06	22.80	<u>38.39</u>	21.87	22.08	<u>23.14</u>	<u>30.94</u>	<u>25.80</u>	24.68	<u>28.38</u>	<b>33.89</b>
Qwen3-VL-8B	81.82	80.52	67.40	64.65	26.04	28.57	25.68	36.34	<u>22.60</u>	15.58	19.93	28.51	20.39	20.78	22.13	30.42
Qwen3-VL-8B $\triangleleft$	<b>87.22</b>	<u>83.12</u>	<b>78.89</b>	<b>70.61</b>	27.03	29.87	23.65	37.56	22.11	18.18	19.93	29.88	24.57	25.97	25.00	33.00
<i>Closed-Source MLLMs</i>																
GPT-4o-mini	82.80	83.12	67.91	67.57	25.06	38.96	27.03	35.99	28.75	19.48	23.48	27.30	-	-	-	-
GPT-4o	82.06	88.31	72.13	69.03	32.43	51.95	29.73	39.07	27.76	20.78	26.01	31.26	25.31	23.38	21.96	28.43
Qwen3-VL-Plus	90.91	88.31	78.14	76.68	28.57	32.47	27.07	43.69	18.43	20.78	25.21	33.74	24.57	24.68	28.21	37.37
Qwen3-VL-Plus $\triangleleft$	92.38	89.61	84.24	76.22	31.77	38.96	29.95	42.41	26.85	18.18	25.93	35.34	24.57	18.18	25.17	36.92
Gemini-2.5-Falsh	85.01	88.31	72.64	73.04	30.47	31.93	45.58	43.57	28.26	23.38	30.07	36.62	26.04	24.68	29.39	37.44
Gemini-2.5-Falsh $\triangleleft$	92.38	93.51	83.11	79.95	34.64	51.95	35.47	51.62	30.47	33.77	31.93	42.28	32.43	32.47	36.32	42.63
Gemini-2.5-Pro $\triangleleft$	<u>93.37</u>	<u>94.81</u>	<u>92.23</u>	<u>84.79</u>	<u>50.37</u>	<u>74.03</u>	<u>53.21</u>	<u>61.13</u>	<u>35.38</u>	<b>44.16</b>	<u>46.11</u>	<u>48.32</u>	<u>36.61</u>	<u>35.06</u>	<u>39.02</u>	<u>48.35</u>
GPT-5 $\triangleleft$	<b>96.81</b>	<b>97.40</b>	<b>97.13</b>	<b>89.64</b>	<b>59.21</b>	<b>81.82</b>	<b>59.46</b>	<b>66.96</b>	<b>44.72</b>	<b>42.86</b>	<b>46.96</b>	<b>53.42</b>	<b>46.93</b>	<b>53.25</b>	<b>53.55</b>	<b>56.72</b>
Random	25.35	27.27	24.69	25.23	24.57	24.67	24.49	24.73	25.79	23.98	24.92	25.03	24.57	24.67	24.76	25.04
Human	98.77	97.40	96.96	96.93	98.77	98.70	98.31	97.39	99.26	98.70	97.47	97.39	99.26	98.70	97.47	97.39

Table 1: Performance of MLLMs on CNSL-bench. AW, FS, and MA indicate air-writing, finger-spelling, and manual-alphabet.  $\triangleleft$  denotes inference with slow thinking. The best result is **bolded**, and the second is underlined.

manual-alphabet, enabling dedicated subset evaluation for sign-linguistical sensitive analysis. Moreover, a sign entry may comprise multiple atomic gestures due to sequential articulation or multi-part realizations, with up to 7 gestures in the most complex cases. The detailed statistics and breakdowns are provided in Appendix A.4.

Overall, CNSL-bench comprises 20,121 questions spanning text, image, and video modalities, with each question grounded in a standardized lexical entry and aligned multimodal evidence. This construction supports systematic and fine-grained evaluation of sign language understanding under a unified benchmark setting.

### 3 Experiments

#### 3.1 Experimental Settings

We evaluate 21 up-to-date MLLMs, spanning open-source families (LLaVA-NeXT, Qwen-VL, InternVL-3.5, GLM-4.1V) and closed-source models (Qwen-Plus/Max, Gemini-2.5, GPT-4/5). Detailed inference configurations and the human evaluation protocol are provided in Appendix B.1.

#### 3.2 Main Results

The overall performance and subcategory comparisons (human vs. representative MLLMs) on CNSL-Bench can be quickly glanced at Figure 3. Table 1 details the overall performance of a wide range of MLLMs on CNSL-bench across modalities (i.e., text, image, and sign video) and manual articulatory forms (i.e., AW, FS, and MA).

**Human-MLLMs performance gap.** Despite recent progress, current MLLMs remain markedly inferior to human-level sign language understanding across all modalities. The strongest proprietary model, GPT-5, achieves overall accuracies of 89.64%, 66.96%, and 56.72% on text, image, and sign language video understanding, respectively. Although GPT-5 substantially outperforms other advanced models, a clear and persistent gap remains when compared to human performance, which consistently reaches approximately 97% across all three modalities. This discrepancy highlights the fundamental difficulty of achieving robust, human-level comprehension of sign language, particularly in visually grounded and temporally

Model	Text				Image				Video <sup>2 fps</sup>				Video <sup>10 fps</sup>			
	AW	FS	MA	All	AW	FS	MA	All	AW	FS	MA	All	AW	FS	MA	All
<i>Fast Thinking</i>																
Qwen3-VL-2B	66.58	74.03	52.70	57.97	23.59	27.27	20.78	32.80	20.39	16.88	17.74	26.38	22.36	23.38	18.07	28.95
Qwen3-VL-8B	81.82	80.52	67.40	64.65	26.04	28.57	25.68	36.34	22.60	15.58	19.93	28.51	20.39	20.78	22.13	30.42
Qwen3-VL-Plus	90.91	88.31	78.14	76.68	28.57	32.47	27.07	43.69	18.43	20.78	25.21	33.74	24.57	24.68	28.21	37.37
Gemini-2.5-Flash	85.01	88.31	72.64	73.04	30.47	31.93	45.58	43.57	28.26	23.38	30.07	36.62	26.04	24.68	29.39	37.44
<i>Slow Thinking</i>																
Qwen3-VL-2B	72.97	75.32	56.25	61.29	26.54	41.56	21.28	35.37	16.71	15.58	19.76	27.76	22.36	27.27	20.95	30.68
Qwen3-VL-8B	87.22	83.12	78.89	70.61	27.03	29.87	23.65	37.56	22.11	18.18	19.93	29.88	24.57	25.97	25.00	33.00
Qwen3-VL-Plus	92.38	89.61	84.24	76.22	31.77	38.96	29.95	42.41	26.85	18.18	25.93	35.34	24.57	18.18	25.17	36.92
Gemini-2.5-Flash	92.38	93.51	83.11	79.95	34.64	51.95	35.47	51.62	30.47	33.77	31.93	42.28	32.43	32.47	36.32	42.63
Gemini-2.5-Pro (L)	91.15	96.10	86.15	81.32	46.68	72.73	52.20	58.09	37.59	42.86	44.76	48.83	34.64	42.86	41.72	47.96
Gemini-2.5-Pro (M)	93.37	94.81	92.23	84.79	50.37	74.03	53.21	61.13	35.38	44.16	46.11	48.32	36.61	35.06	39.02	48.35
Gemini-2.5-Pro (H)	92.63	94.81	90.71	84.84	49.63	74.03	54.05	61.92	36.61	38.96	42.40	48.17	38.57	36.36	45.10	48.59
GPT-5 (L)	97.05	97.40	95.61	88.94	55.53	74.03	60.64	66.77	40.54	38.96	46.45	51.89	41.03	48.05	45.10	53.13
GPT-5 (M)	96.81	97.40	97.13	89.64	59.21	81.82	59.46	66.96	44.72	42.86	46.96	53.42	46.93	53.25	53.55	56.72
GPT-5 (H)	97.05	98.70	96.62	89.95	63.14	83.12	59.46	68.34	42.51	37.66	48.99	53.09	44.58	38.96	46.45	54.01

Table 2: Numerical results with test-time scaling on reasoning models. L, M, H: low, medium, and high reasoning effort on the process of thinking before generating an answer.

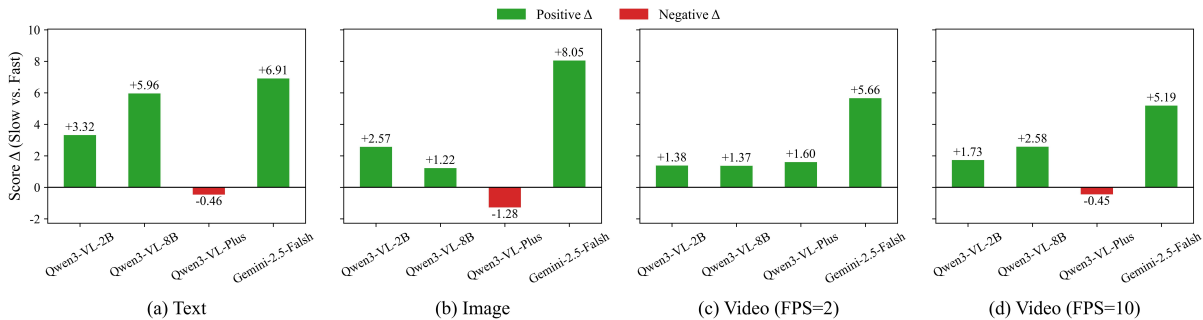


Figure 4: CoT gains across models and multimodal subsets.

complex settings.

### Modality-dependent performance imbalance.

A pronounced modality imbalance is consistently observed in MLLMs’ sign language understanding. Across nearly all models, performance is highest for textual descriptions, while accuracy drops substantially for illustrative images and further degrades for sign language videos. This trend holds for both open-source and closed-source systems and becomes more pronounced in the video setting. These results indicate that, although text-centric language modeling is relatively well developed, robust visual grounding and temporal modeling remain major challenges for current MLLMs.

### Uneven understanding across manual articulatory forms.

MLLMs exhibit uneven comprehension across different manual articulatory forms. As shown in Table 1, models consistently achieve stronger performance on finger-spelling than on air-writing and manual-alphabet across all modalities. This disparity suggests that current models handle

more discrete and character-like sign components more reliably, while continuous, shape-intensive, or motion-dependent articulations remain difficult.

### Narrowing gap between open- and closed-source MLLMs.

The performance gap between open-source and proprietary MLLMs is rapidly narrowing. Several small-scale open-source models, including GLM-4.1V-9B, InternVL-3.5-8B, and Qwen3-VL-8B, attain performance comparable to lightweight commercial systems such as GPT-4o-mini and Gemini-2.5-Flash across multiple modalities. These models even surpass proprietary counterparts on specific subsets, underscoring the rapid advancement and increasing competitiveness of open-source MLLMs in sign language understanding. Nevertheless, despite these encouraging trends, open-source models still need to make continued progress on challenging subsets and to match higher-capacity proprietary systems, particularly under multimodal settings.

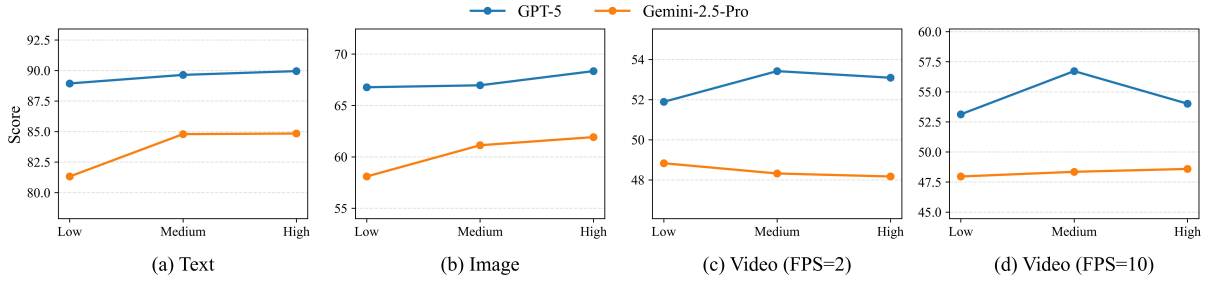


Figure 5: Top-tier reasoning models, i.e., GPT-5 and Gemini-2.5-Pro, exhibit a boundary effect in test-time scaling as reasoning effort increases from low to high, especially in video input settings.

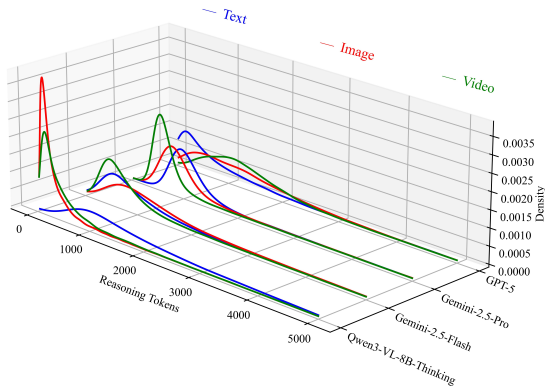


Figure 6: KDE plot of reasoning tokens. Videos are at 10 frames per second. Zoom in for better visualization.

### 3.3 Additional Analysis

**Test-Time Scaling.** To better contextualize empirical findings from the main results, we examine test-time scaling (reasoning) via CoT on MLLMs with explicit reasoning capabilities. We compare fast-thinking and slow-thinking settings across modalities and manual articulatory forms to assess when additional reasoning is beneficial for sign language understanding. After carefully analyzing these results, we draw several conclusions:

(1) *CoT gains vary substantially across models.* As shown in Table 2 and Figure 4, CoT gains vary substantially across models and subsets. Gemini-2.5-Flash consistently benefits most from slow thinking, achieving an average improvement of 6.45%, whereas Qwen3-VL-Plus exhibits negative gains on several subsets, including textual descriptions, illustrative images, and sign language videos (with 10 frames per second). The repeated evaluations in Qwen3-VL-Plus yield consistent negative results, suggesting that reasoning does not universally improve performance and may be sensitive to model-specific inference behavior.

(2) *CoT gains plateau for top-tier models.* Fig-

ure 5 shows that for top-tier models such as Gemini-2.5-Pro and GPT-5, increasing reasoning effort from low to high yields marginal or no gains, and can even degrade performance on sign language video inputs. This indicates a boundary effect of test-time scaling, where additional reasoning tokens no longer translate into improved understanding once a strong baseline is reached.

(3) *Reasoning quality differs by modality.* Figure 6 reveals that models exhibit large variance in reasoning length and modality preference. The reasoning tokens vary by nearly an order of magnitude across models under textual input. The Qwen3-VL and Gemini-2.5-Pro tend to allocate significantly more reasoning effort to text than to image or video inputs. Such modality-dependent behavior indicates incomplete multimodal alignment, which is especially problematic given the inherently visual-language nature of sign language.

(4) *Reasoning length correlates with task difficulty.* MLLMs tend to generate longer reasoning chains for incorrectly answered instances. The ratio between CoT length for incorrect versus correct predictions of GPT-5-M with text input reaches up to 2.89. This pattern is consistent across most models and modalities, suggesting that models spend more effort on harder cases, partially mirroring human problem-solving behavior. Please see Appendix C.1 for detailed analysis.

These results indicate that CoT primarily serves as a diagnostic lens rather than a universal performance enhancer, revealing fundamental challenges in multimodal reasoning and sign language-specific understanding that persist in MLLMs.

**Prompt Tokens & Instruction Following.** Beyond reasoning-oriented analyses, we further examine prompt token consumption and instruction-following robustness as complementary diagnostic factors for sign language understanding in MLLMs.



Figure 7: Video-based sign language understanding failure for “laptop”. Despite correctly recognizing the typing motion in a continuous sign language video, models fail to recover the intended meaning, indicating that temporal visual recognition alone is insufficient for reliable sign language understanding.

Prompt token usage varies substantially across input modalities and model families, with image and video inputs incurring orders-of-magnitude higher token consumption than text, potentially constraining effective context and reasoning budgets. In addition, instruction-following robustness differs markedly across models and modalities: while most large-scale models remain stable, several smaller-capacity models and certain MLLM families exhibit pronounced failures, particularly on sign language videos. Notably, explicit reasoning mechanisms may also interact non-trivially with instruction adherence. Due to the space limitation, the detailed quantitative analyses are provided in Appendix C.2 and Appendix C.3, respectively.

### 3.4 Case Studies

To provide intuitive insights into the sources of observed performance differences and further illustrate the limitations revealed by the CNSL-benchmark, we present a case study in Figure 7. A primary observation is the pronounced modality sensitivity. For the concept *laptop*, the textual description explicitly mentions “striking keys,” allowing models to easily deduce the correct answer. However, under video inputs, even advanced

models (e.g., Gemini-2.5-Pro, Qwen3-VL) fail to ground the visual motion correctly. Instead of recognizing the “typing” and “opening” gestures, they hallucinate unrelated motions (e.g., “crossing” gestures in Qwen3-VL-8B or “foundation” gestures in Gemini-2.5-Pro) and consistently misclassify the sign as “Because” (Option C). This contrast highlights that while MLLMs possess strong textual reasoning, their ability to parse complex temporal visual cues in sign language remains fragile. Beyond modality effects, our analysis reveals challenges in implicit semantic association (e.g., “smell” vs “air”) and emerging symbolic understanding (e.g., correctly mapping visual signs to Chinese characters), see cases in Appendix C.4.

## 4 Related Works

### 4.1 Sign Language Understanding

Over the past few decades, the sign language research community has primarily focused on sign language recognition (SLR), emphasizing the identification of gloss-level units<sup>2</sup> in isolated words (Joze and Koller, 2019; Li et al., 2020)

<sup>2</sup>Glosses are spoken-language textual units that approximately capture the meaning of sign language.

or in continuous sequences (Koller et al., 2015; Huang et al., 2018). The increasing availability of large-scale sign language translation datasets (Camgoz et al., 2018; Zhou et al., 2021; Duarte et al., 2021; Tanzer and Zhang, 2024) has further driven progress in end-to-end sign language translation (SLT) (Camgoz et al., 2020; Chen et al., 2022; Fu et al., 2024; Zhao et al., 2024; Zhang et al., 2025; Fu et al., 2025a). In parallel, large language models (LLMs) have become general-NLP purpose backbones for a wide range of NLP tasks (Brown et al., 2020; Touvron et al., 2023), motivating recent efforts to incorporate them into sign language research for their strong language modeling and generation capabilities. In particular, prior work has leveraged LLMs either as enhanced text decoders (Wong et al., 2023; Gong et al., 2024; Liu et al., 2024) or as semantic enhancement modules to improve SLT systems (Guo et al., 2025; Kim et al., 2025; Liu et al., 2025; Jang et al., 2025).

Existing work in this line is largely centered on task- or dataset-specific adaptation of LLMs within sign language pipelines. In contrast, systematic evaluation of models’ intrinsic sign language understanding, particularly for MLLMs operating directly on images and videos, has received comparatively less attention. Different from focusing on specific downstream tasks or datasets, we propose CNSL-bench, a comprehensive Chinese National Sign Language benchmark designed for evaluating MLLMs in sign language understanding.

## 4.2 MLLM benchmarks

Recent progress in multimodal large language models (MLLMs) has coincided with the establishment of standardized benchmarks designed to evaluate multimodal understanding. Early efforts primarily focused on image-based evaluation, including visual question answering, caption-based reasoning, and broader vision-language understanding benchmarks that assess perception, grounding, and semantic reasoning over static images (Fu et al., 2025b; Yue et al., 2024; Li et al., 2024; Chen et al., 2024a). More recently, the community has extended benchmarking to video-centric settings, introducing datasets and evaluation protocols that emphasize temporal grounding, event understanding, long-context reasoning, and multimodal dialogue over videos (Maaz et al., 2024; Zhou et al., 2025; Fu et al., 2025c).

Despite their broad coverage, most existing MLLM benchmarks are designed for general-

domain image and video understanding, where the visual content and semantics are dominated by everyday objects, scenes, actions, and events. Benchmarks that explicitly target sign language remain relatively scarce, despite the need for fine-grained modeling of hand articulation, motion trajectories, and linguistically grounded semantics. In contrast, CNSL-bench is constructed as a dedicated evaluation benchmark for sign language understanding, enabling systematic assessment of MLLMs under aligned textual descriptions, illustrative images, and sign language videos.

## 5 Conclusion

In this work, we introduce CNSL-bench, a multimodal benchmark centered on sign language for evaluating sign language understanding in MLLMs, with authoritative grounding in officially standardized sign language resources. Through extensive evaluation of a variety of open- and closed-source models, we systematically uncover persistent challenges in current MLLMs’ ability to comprehend sign language across modalities and manual articulatory forms. We anticipate that CNSL-bench will serve as both a diagnostic foundation and a reference resource for future research toward more robust, reliable, and human-aligned MLLMs.

## Limitations

CNSL-bench focuses on lexical-level canonical sign understanding and adopts a multiple-choice formulation to enable controlled, scalable, and reproducible evaluation across modalities. While this design does not directly assess open-ended sign language generation, it is motivated by the observation that current MLLMs remain unreliable in interpreting free-form sign language. This limitation highlights a fundamental gap between existing model capacities and the demands of open-ended sign interpretation, underscoring the necessity of establishing a diagnostic benchmark target at sign language understanding. In addition, CNSL-bench is centered on Chinese National Sign Language and emphasizes authoritative semantic grounding over linguistic breadth, and thus does not cover cross-linguistic, regional, or dialectal variation present in other sign languages. Extending evaluation to multilingual and multi-regional sign languages, as well as to more open-ended and compositional settings, remains an important direction for future work.

## Ethical Considerations

**Data Access.** CNSL-bench is constructed by aligning publicly available resources with recently released sign language video datasets and does not involve new data collection or additional human participants. The textual descriptions and illustrative images are sourced from officially published materials released by the Ministry of Education of the People’s Republic of China and are publicly accessible for educational and general communication purposes. The sign language videos are derived from an open-source dataset whose data collection and public release were approved by the Ethical Review Board of Leshan Normal University, with all participants providing informed consent for the use and publication of their identity information and recordings (Ethical Review Number: LSNU-KYLL2025-02-15) (Jin et al., 2025).

**Human Participant.** Our benchmark involves the human assessment, and we invited a professional team consisting of one professor specializing in sign language linguistics and three sign-language students (including one hearing-impaired student). Each student has at least one year of classroom studying experience in sign language; their instructors include the invited professor and Deaf sign language teachers from a local special education institute. Participants are compensated with \$30 to complete each task (about two hours of work). Overall, one expert and three students are engaged to fulfill the human assessment tasks.

## Acknowledgements

We are grateful for the efforts and time of the reviewers and the committee. This work was supported in part by the National Natural Science Foundation of China under Grant 62476232, Grant 62076211, and in part by First Batch of Projects for the 2025 “Intergovernmental International Science, Technology and Innovation Cooperation” of the National Key Research and Development Program of China under Grant 2025YFE0121700.

## References

Sobhan Asasi, Mohamed Ilyas Lakhali, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. [Beyond gloss: A hand-centric framework for gloss-free sign language translation](#). *Preprint*, arXiv:2507.23575.

Katherine Atwell, Danielle Bragg, and Malihe Alikhani. 2024. [Studying and mitigating biases in sign lan-](#)

[guage understanding models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 268–283, Miami, Florida, USA. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. [Multi-channel transformers for multi-articulatory sign language translation](#). In *Computer Vision – ECCV 2020 Workshops, Lecture Notes in Computer Science*, pages 301–319, Cham. Springer International Publishing.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. [M<sup>3</sup>cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. [A simple multi-modality transfer learning baseline for sign language translation](#). In

- 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5110–5120.
- Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024b. 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*, pages 7071–7081, Torino, Italia. ELRA and ICCL.
- China Disabled Persons' Federation, China Association of Persons with Hearing Disabilities, and National Center for Sign Language and Braille. 2019. *Lexicon of Expressions in Chinese National Sign Language*. Huaxia Publishing House, Beijing, China.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasapat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. Preprint, arXiv:2507.06261.
- Committee for Language Reform of China. 1957. *Scheme of the chinese phonetic alphabet*. In *The 5th Session of The 1st National People's Congress*, Beijing, China.
- Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language ai research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia. ELRA and ICCL.
- Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i-Nieto. 2021. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2735–2744.
- Biao Fu, Liang Zhang, Peigen Ye, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2025a. *Improving end-to-end sign language translation via multi-level contrastive learning*. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1230–1242.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025b. *Mme: A comprehensive evaluation benchmark for multimodal large language models*. Preprint, arXiv:2306.13394.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025c. *Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis*. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118.
- Honghao Fu, Liang Zhang, Biao Fu, Rui Zhao, Jinsong Su, Xiaodong Shi, and Yidong Chen. 2024. *Signer diversity-driven data augmentation for signer-independent sign language translation*. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2182–2193, Mexico City, Mexico. Association for Computational Linguistics.
- Team GLM-V, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. *Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning*. Preprint, arXiv:2507.01006.
- Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372.
- Jianyuan Guo, Peike Li, and Trevor Cohn. 2025. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.
- Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18*, pages 2257–2264, New Orleans, Louisiana, USA. AAAI Press.
- 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 (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.
- Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8742–8752.

- Peng Jin, Hongkai Li, Jun Yang, Yazhou Ren, Yuhao Li, Lilan Zhou, Jin Liu, Mei Zhang, Xiaorong Pu, and Siyuan Jing. 2025. [A large dataset covering the chinese national sign language for dual-view isolated sign language recognition](#). *Scientific Data*, 12(1):660.
- Hamid Reza Vaezi Joze and Oscar Koller. 2019. [Ms-asl: A large-scale data set and benchmark for understanding american sign language](#). *Preprint*, arXiv:1812.01053.
- Jungeun Kim, Hyeonwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. [Leveraging the power of mllms for gloss-free sign language translation](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21048–21058.
- Oscar Koller, Jens Forster, and Hermann Ney. 2015. [Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers](#). *Computer Vision and Image Understanding*, 141:108–125.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. [Seed-bench: Benchmarking multimodal large language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.
- Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. [Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison](#). In *2020 IEEE Winter Conference on Applications of Computer Vision*, pages 1448–1458.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, pages 34892–34916, Red Hook, NY, USA. Curran Associates Inc.
- Yuqi Liu, Wenqian Zhang, Sihan Ren, Chengyu Huang, Jingyi Yu, and Lan Xu. 2025. [Scope: Sign language contextual processing with embedding from llms](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5739–5747.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. [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)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.
- Ministry of Education of the People’s Republic of China, State Language Commission, and China Disabled Persons’ Federation. 2018a. *Chinese manual alphabet*. Huaxia Publishing House, Beijing, China.
- Ministry of Education of the People’s Republic of China, State Language Commission, and China Disabled Persons’ Federation. 2018b. *Lexicon of Common Expressions in Chinese National Sign Language*. Huaxia Publishing House, Beijing, China.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Zhi Rao, Yucheng Zhou, Benjia Zhou, Yiqing Huang, Sergio Escalera, and Jun Wan. 2025. [Rvlf: A reinforcing vision-language framework for gloss-free sign language translation](#). *Preprint*, arXiv:2512.07273.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. [Vlm-r1: A stable and generalizable r1-style large vision-language model](#). *Preprint*, arXiv:2504.07615.
- Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2021. [Fingerspelling detection in american sign language](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4166–4175.
- Garrett Tanzer and Biao Zhang. 2024. [Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus](#). In *The Thirteenth International Conference on Learning Representations*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. [Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *Preprint*, arXiv:2508.18265.

- Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2023. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.
- James C. Woodward. 1972. Implications for sociolinguistic research among the deaf. *Sign Language Studies*, 1(1):1–7.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Ruiquan Zhang, Rui Zhao, Zhicong Wu, Liang Zhang, Haoqi Zhang, and Yidong Chen. 2025. Dynamic feature fusion for sign language translation using hypernetworks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6227–6239, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multi-modal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.
- Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. [Conditional variational autoencoder for sign language translation with cross-modal alignment](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19643–19651.
- Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. [Improving sign language translation with monolingual data by sign back-translation](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. [Mlvu: Benchmarking multi-task long video understanding](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701.

## A CNSL-bench

### A.1 Dataset Alignment

This section details the data processing and alignment procedures underlying the construction of CNSL-bench. The textual descriptions and illustrative images are sourced from the National Common Sign Language Dictionary, which contains 8,214 commonly used sign entries and serves as an authoritative reference for standardized CSL (Ministry of Education of the People’s Republic of China et al., 2018b; China Disabled Persons’ Federation et al., 2019). During data preparation, we identify several forms of redundancy in the original dictionary. First, some entries share identical meanings and identical sign realizations, which are merged into a single sign entry. Second, certain lexical items correspond to multiple meanings, which are distinguished in the dictionary using auxiliary index markers. Third, some entries share the same lexical form and meaning but differ in their sign realizations. To ensure a consistent representation, we remove these auxiliary markers and retain all valid lexical variants associated with each sign realization, followed by sign-level processing. As a result, a set of 6,707 unique sign entries is obtained. In addition, the dictionary includes explanatory and instructional content intended for human readers, and we remove such descriptive text and preserve only the core lexical information. Moreover, the dictionary does not explicitly annotate linguistically manual articulatory forms such as air-writing, finger-spelling, or manual-alphabet, we manually identify and annotate these categories to support fine-grained analysis.

The sign language videos are sourced from the CNSL-DP dataset (Jin et al., 2025), which was collected under institutional ethical approval and provides synchronized video recordings for individual sign entries. For each sign, multiple video instances from different signers are available. To ensure both consistency and representativeness, we select one representative recording per sign entry for inclusion in the benchmark. The original videos are recorded at a resolution of  $1920 \times 1080$  and 50 frames per second, with the signer centered in the frame. We uniformly downsample the videos to 24 frames per second and apply center cropping followed by resizing to  $512 \times 512$  to standardize visual inputs. In cases where multiple synonymous lexical entries correspond to the same sign realization, the original CNSL-DP dataset retains

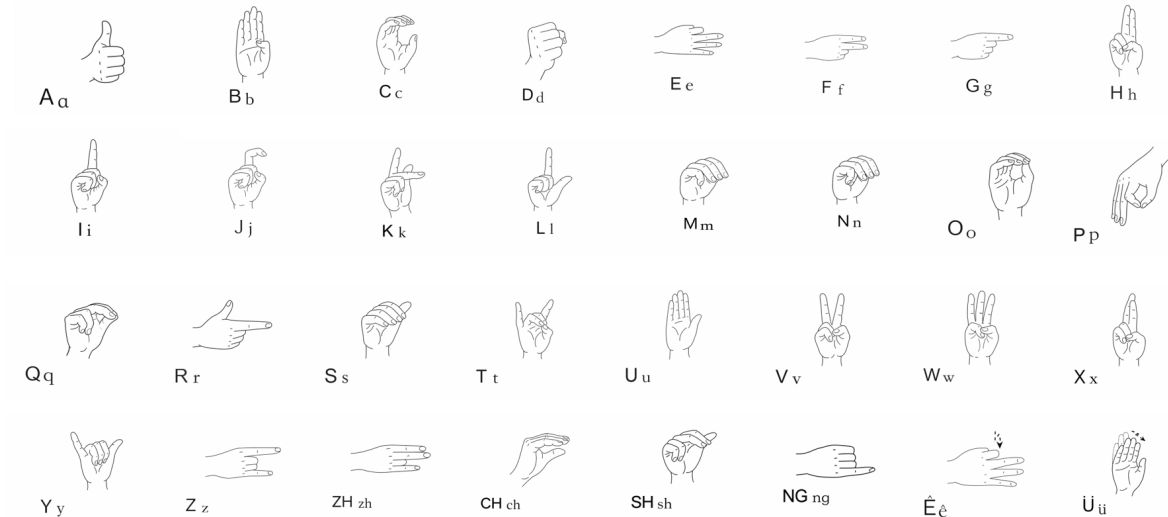


Figure 8: All manual alphabets in the *Chinese Manual Alphabet*, including 26 single-letter alphabets, 4 double-letter alphabets, and 2 alphabets with symbols.

only a single representative form. To construct a lexically complete benchmark aligned with the official dictionary, we explicitly recover the omitted synonymous entries and re-associate them with the corresponding video instances during dataset alignment, thereby restoring the full set of lexical variants for each sign. As a result, CNSL-bench establishes a unified mapping among textual descriptions, illustrative images, and sign language videos, thereby constructing a coherent and well-aligned benchmark for evaluating sign language understanding.

## A.2 The China Manual Alphabet

Figure 8 shows all manual alphabets in the Chinese Manual Alphabet, including 26 single-letter manual alphabets, 4 double-letter manual alphabets, and 2 manual alphabets with symbols.

## A.3 Detailed Task Definition

This section provides additional description supporting the task formulation and option construction strategies adopted in CNSL-bench. We first examine the feasibility of open-ended sign language understanding and find that current state-of-the-art MLLMs remain highly unreliable in this setting. As illustrated in Figure 9, we draw a case from CSL-Daily (Zhou et al., 2021), a Chinese sign language dataset targeted at sign language recognition and translation. The results show that even for a short video expressing a simple and common sentence (e.g., “The weather is very nice today, neither cold nor hot”), flagship models such as

Gemini-3-Pro and GPT-5.1 fail to recover the intended meaning, instead misinterpreting isolated handshapes or hallucinating unrelated lexical concepts. The observed failures suggest that existing models struggle to robustly integrate temporal dynamics, sequential articulation, and lexical composition in free-form generation, leading to unstable and difficult-to-interpret outputs.

Distractor	Text	Image	Video <sup>2</sup>	Video <sup>10</sup>
Random	72.41	38.83	32.06	33.95
Semantic-Based	64.33	36.95	32.60	32.47

Table 3: An analysis of distractor on Qwen-VL-Plus.

In addition, we further analyze different option construction strategies for the multiple-choice formulation. As shown in Table 3, semantics-based distractors yield slightly lower absolute accuracy than random sampling across modalities on a representative model (i.e., Qwen-VL-Plus), yet they lead to qualitatively consistent conclusions about model performance. This indicates that the overall ranking and modality-dependent trends are stable under alternative distractor designs, and that benchmark outcomes are not driven by overly obvious negatives. In other words, random sampling already provides sufficiently challenging and reliable evaluation signals for CNSL-bench. Since semantics-based distractors introduce additional engineering overhead and potential sensitivity to similarity heuristics without changing the main findings, we adopt random option sampling to streamline the



Figure 9: An example of open-ended sign language understanding from the CSL-Daily dataset (Zhou et al., 2021). The input video expresses a simple sentence (“The weather is very nice today, neither cold nor hot”), yet flagship models (e.g., Gemini-3-Pro and GPT-5.1) fail to generate the correct meaning. Note that the realization of concepts such as “weather” in this example differs from the canonical forms used in CNSL-bench, reflecting natural variation in sign expression, while remaining readily interpretable to human signers.

benchmark design and support robust, reproducible evaluation.

#### A.4 Detailed Dataset Statistics

The detailed statistics of CNSL-bench are illustrated in Table 4, and the frame count is calculated at a rate of 24 frames per second.

Subset	#Sign Entry	#Frames
Air-Writing	407	99.8
Finger-Spelling	77	109.2
Manual-Alphabet	592	98.0
w/ 1 gesture	2,977	89.2
w/ 2 gestures	3,287	100.2
w/ 3 gestures	369	120.9
w/ 4 gestures	62	136.4
w/ 5 gestures	8	148.4
w/ 6 gestures	2	161.0
w/ 7 gestures	2	201.5
All	6707	96.89

Table 4: The detailed statistics of CNSL-bench.

## B Experiments

### B.1 Experimental Settings

**MLLMs Participants.** A total of 21 MLLMs (13 Open-source and 8 closed-source MLLMs) are included for validation, which includes a) 3 open&closed-source Image MLLMs: LLaVA-NeXT (Mistral-7B) (Liu et al., 2024), Qwen-VL-Plus/Max (Bai et al., 2023), b) 12 open-source MLLMs: Qwen2-VL-2B/7B (Wang et al., 2024), Qwen2.5-VL-3B/7B (Bai et al., 2025b), Intern3.5-VL-2B/8B (Wang et al., 2025), Qwen3-VL-2B/8B-Instruct, Qwen3-VL-2B/8B-Thinking (Bai et al., 2025a), LLaVA-NeXT-Video-7B (Liu et al., 2024), and GLM-4.1V-9B-Thinking (GLM-V. et al., 2025), and c) 6 close-source MLLMs: Qwen3-VL-Plus (Bai et al., 2025a), Gemini-2.5-Flash, Gemini-2.5-Pro (Comanici et al., 2025), GPT-4o-mini, GPT-4o, and GPT-5 (OpenAI et al., 2024).

**Evaluation Details.** For open source MLLMs, we primarily use the Hugging Face transformers library<sup>3</sup> for model inference. To accelerate decoding for *thinking* models under the slow-thinking setting, we adopt vLLM<sup>4</sup> as the inference backend, and set the maximum generation length to 8,192 tokens per response. For closed-source MLLMs,

<sup>3</sup><https://hugging-face.cn/docs/transformers>

<sup>4</sup><https://github.com/vllm-project/vllm>

any samples that fail due to API errors, timeouts, or malformed generations are excluded from scoring, so that the reported metric scores are computed only over valid outputs. As for video inputs, we evaluate two frame sampling rates: 2 fps (the default in many video-understanding benchmarks) and a denser 10 fps setting, which better captures the high-speed and fine-grained spatiotemporal motions characteristic of sign language.

To mitigate input length constraints, when a dense sampling rate (e.g., FPS=10) would exceed a model’s maximum context window, we adaptively resample the video and include as many frames as possible *without* surpassing the input length limit (e.g., for LLaVA-NeXT-Video, we reduce the sampling rate but maximize the number of frames allowed within its context budget). For the GPT-series models, we uniformly cap the visual input to at most 50 frames to comply with the API restriction. Unless stated otherwise, all hyperparameters follow the official recommendations for each model, as summarized in Table 6. Accuracy is computed by an exact match between the model prediction and the ground-truth answer. For thinking models, we manually extract the final answer from the generated response to avoid conflating intermediate reasoning with the predicted label.

**Human Assessment.** Deaf community involvement is essential for developing sign language understanding systems (Yin et al., 2021; Atwell et al., 2024). To establish a human reference for CNSL-bench, we invited a professional team consisting of one professor specializing in sign language linguistics and three sign-language students (including one hearing-impaired student). Each student has at least one year of classroom studying experience in sign language; their instructors include the invited professor and Deaf sign language teachers from a local special education institute. To make the evaluation feasible while preserving articulation diversity, we constructed a 1,500-entry subset from the 6,707 sign entries. Specifically, we retained all entries involving air-writing, finger-spelling, and manual-alphabet articulations, yielding 1,018 entries, and then randomly sampled an additional 482 entries from the remaining gesture-only entries. For each entry, we generated three multiple-choice questions corresponding to the aligned textual description, illustrative image, and sign language video, resulting in 4,500 questions in total. Each evaluator completed all questions, and we

report the average accuracy over the three student evaluators as the human performance.

## C Additional Analysis

### C.1 Detailed Reasoning Tokens

Table 5 reports the average reasoning tokens generated by each model, stratified by correctness and modality. Across all evaluated MLLMs, incorrect predictions are consistently associated with substantially longer reasoning traces than correct ones, reinforcing the observation that models tend to “think longer” when facing harder or ambiguous inputs, partially mirroring human problem-solving behavior. This effect is particularly pronounced for stronger models. For instance, GPT-5 (M) exhibits a ratio of 2.89 between incorrect and correct cases under text input, indicating that failed attempts often trigger nearly three times as many reasoning tokens. Similar trends are observed in Gemini-2.5-Flash, whose ratios exceed 2.0 in the text modality.

This phenomenon generalizes to image and video inputs, albeit with a noticeably attenuated magnitude. In image settings, the ratios between incorrect and correct reasoning length typically fall within 1.2–1.7, while with video input, the ratios further decrease to around 1.0–1.3. This compression suggests that the tendency to engage in longer reasoning on more difficult cases, which is clearly observed in the text-only setting, becomes less pronounced once multimodal perception is introduced. Rather than reflecting increased confidence, the reduced gap more plausibly indicates that multimodal perception and alignment imperfections constrain the model’s ability to adaptively allocate reasoning effort: when visual evidence is noisy, under-specified, or imperfectly aligned with the language space, the model may fail to trigger longer, exploratory reasoning even on genuinely difficult instances. A further supporting signal is that increasing the temporal resolution of video inputs does not yield a systematic restoration of the gap. The ratios observed at 2 FPS and 10 FPS remain highly similar across models, including GPT-5 and Gemini-2.5 variants, despite the substantially increased number of frames. This suggests that the attenuation is not primarily driven by insufficient temporal evidence, but rather by broader limitations in multimodal understanding, such as imperfect robustness in visual feature extraction, temporal integration, or cross-modal grounding, which prevent additional frames from translating into more accurately calibrated

Model	Text				Image				Video <sup>2 fps</sup>				Video <sup>10 fps</sup>			
	Correct	Fault	All	Ratio	Correct	Fault	All	Ratio	Correct	Fault	All	Ratio	Correct	Fault	All	Ratio
Qwen3-VL-8B	1,688	2,905	2,046	1.72	255	303	285	1.19	390	414	407	1.06	252	275	267	1.09
GLM-4.1V-9B	179	304	218	1.70	278	379	339	1.36	335	400	382	1.19	318	397	374	1.25
Qwen3-VL-Plus	1,212	2,124	1,429	1.75	360	431	401	1.20	1,264	1,411	1,359	1.12	322	317	319	0.98
Gemini-2.5-Flash	818	1,755	1,006	2.15	1,090	1,827	1,447	1.67	719	938	845	1.30	720	934	843	1.30
Gemini-2.5-Pro (L)	347	349	347	1.00	132	135	133	1.02	76	73	74	0.95	82	84	83	1.02
Gemini-2.5-Pro (M)	1,133	1,755	1,228	1.55	941	1,281	1,073	1.36	700	788	746	1.13	708	803	757	1.14
Gemini-2.5-Pro (H)	365	532	390	1.46	915	1,177	1,015	1.29	675	757	718	1.12	699	797	749	1.14
GPT-5 (L)	248	637	291	2.57	339	494	391	1.45	325	395	359	1.22	353	442	394	1.25
GPT-5 (M)	635	1,837	759	2.89	1,017	1,614	1,214	1.59	1,093	1,420	1,245	1.30	1,312	1,701	1,480	1.30
GPT-5 (H)	1,392	3,737	1,628	2.68	2,248	3,211	2,553	1.43	2,305	2,777	2,527	1.20	2,597	3,150	2,852	1.21

Table 5: Detailed reasoning tokens across models and modalities. L, M, H: low, medium, and high reasoning effort on the process of thinking before generating an answer.

Model	Text			V-L		
	top_p	top_k	T	top_p	top_k	T
LLaVA-NeXT-7B	0.95	50	1.0	0.95	50	1.0
LLaVA-NeXT-Video-7B	0.95	50	1.0	0.95	50	1.0
Qwen2/2.5-VL-Instruct	0.95	50	1.0	0.95	50	1.0
Intern-VL-3.5	0.95	50	1.0	0.95	50	1.0
GLM-4.1V-9B	0.95	50	1.0	0.95	50	1.0
Qwen3-VL-Instruct	1.00	40	1.0	0.80	20	0.7
Qwen3-VL	0.95	20	1.0	0.95	20	1.0
Gemini-2.5-Series	0.95	-	1.0	0.95	-	1.0
GPT-Series	1.0	-	1.0	1.0	-	1.0

Table 6: Detailed hyperparameters. T means temperature. V-L denotes multimodal input settings.

reasoning effort.

Conclusively, the results point to a modality-dependent divergence in test-time behavior specific to sign language understanding. Although MLLMs display difficulty-sensitive processing in text-based settings, this characteristic is notably attenuated for visual inputs. Such attenuation suggests that current models may struggle to effectively utilize visual linguistic cues, indicating that limitations in multimodal perception and alignment contribute to the reduced adaptability observed in sign language understanding tasks.

## C.2 Detailed Prompt Tokens

As shown in Table 8, the consumption of prompt tokens varies substantially across both input modalities and models. For a fixed model, image and video inputs consistently generate orders of magnitude more prompt tokens than text, resulting in an extreme length gap in multimodal processing. This disparity in multimodal processing may partially account for the observed performance discrepancies across modalities. Beyond modality effects, the token usage also differs markedly under identical text inputs, which can be largely attributed to heterogeneous tokenization and visual encoding

strategies (e.g., LLaVA-NeXT vs. Qwen). Such tokenizer-induced prompt length variations may further affect effective context allocation and reasoning budget, introducing an additional source of performance variability in cross-model comparisons. Finally, closed-source models exhibit distinct budget characteristics shaped by their pricing-oriented design choices. Although GPT-4o-mini offers a lower per-token cost, its substantially higher token consumption for multimodal inputs results in significantly increased overall usage, leading us to exclude it from further evaluation due to prohibitive cost considerations.

## C.3 Instruction Following

As shown in Table 7, instruction adherence varies substantially across models and input modalities. While most large-scale open-source and proprietary MLLMs achieve near-perfect instruction-following accuracy across settings, several smaller-capacity models and certain MLLM families exhibit pronounced failures. Specifically, InternVL-3.5-2B maintains high accuracy on text and image inputs but collapses to around 15% accuracy on sign language videos, indicating severe difficulty in jointly satisfying visual, temporal, and task-level constraints. In contrast, an opposite pattern is observed in Qwen2-VL-2B and models from the LLaVA-Next family, where instruction-following performance is already unstable across different modalities. Beyond model scale, we further observe that explicit CoT mechanisms may interact negatively with instruction adherence, likely due to verbosity and drifting constraints. For example, compared to Qwen3-VL-8B-Instruct, Qwen3-VL-8B-Thinking (marked with  $\not\rightarrow$  in Table 7) exhibits a slight but consistent degradation in instruction-following accuracy. A similar trend is observed in Gemini-2.5-Flash, whose instruction-following performance

Model	Text				Image				Video <sup>2</sup> fps				Video <sup>10</sup> fps			
	AW	FS	MA	All	AW	FS	MA	All	AW	FS	MA	All	AW	FS	MA	All
<i>Open&amp; Close -source Image MLLMs</i>																
LLaVA-NeXT-7B	1.2	6.5	2.5	4.9	61.2	54.6	62.3	62.8	68.8	71.4	73.7	71.2	-	-	-	-
Qwen-VL-Plus	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.8	100.0	99.9
Qwen-VL-Max	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
<i>Open-Source MLLMs</i>																
Qwen2-VL-2B	85.5	83.1	85.0	83.4	98.5	97.4	99.0	98.3	100.0	100.0	100.0	99.9	97.1	90.9	98.3	97.7
Qwen2.5-VL-3B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intern-VL-3.5-2B	99.8	100.0	99.7	99.4	99.3	97.4	97.8	97.5	15.2	15.6	16.4	15.3	17.4	14.3	16.4	15.4
Qwen3-VL-2B	99.8	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
LLaVA-NeXT-Video-7B	3.2	3.9	3.7	4.8	45.2	44.2	47.1	49.8	63.6	58.4	63.7	64.0	60.0	58.4	63.3	62.6
Qwen2-VL-7B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen2.5-VL-7B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
GLM-4.1V-9B <sup>⚡</sup>	99.5	100.0	99.7	99.8	98.8	98.7	99.3	99.1	99.5	100.0	99.2	99.5	99.3	98.7	99.2	99.4
Intern-VL-3.5-8B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen3-VL-8B-Instruct	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen3-VL-8B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen3-VL-8B <sup>⚡</sup>	99.0	96.1	94.9	99.1	100.0	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0
<i>Closed-Source MLLMs</i>																
Qwen3-VL-Plus <sup>⚡</sup>	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Gemini-2.5-Flash	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Gemini-2.5-Flash <sup>⚡</sup>	99.8	100.0	99.3	99.8	92.9	92.2	92.9	96.3	98.3	96.1	97.5	98.4	97.5	98.7	99.0	98.4
Gemini-2.5-Pro <sup>⚡</sup>	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
GPT-4o-mini	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	-	-	-
GPT-4o	99.8	98.7	99.8	99.8	98.8	100.0	99.2	98.9	99.5	100.0	99.3	99.2	99.0	98.7	98.3	98.6
GPT-5 <sup>⚡</sup>	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 7: Instruction Following. <sup>⚡</sup> denotes inference with slow thinking.

Model	Text	Image	Video <sup>2</sup>	Video <sup>10</sup>
<i>Open-Source MLLMs</i>				
LLaVA-NeXT-7B	272	2,453	8,913	-
LLaVA-NeXT-Video <sup>⊗</sup>	307	2,480	1,306	3,698
Qwen2/2.5-VL	169	742	1,413	6,669
Intern-VL-3.5	195	415	2,102	10,659
Qwen3-VL	158	590	1,163	5,445
GLM-4.1V-9B <sup>⚡</sup>	153	726	1,409	6,713
<i>Closed-Source MLLMs</i>				
Qwen-VL-Plus	218	1,519	3,552	15,586
Qwen-VL-Max	218	1,621	3,673	16,892
Qwen3-VL-Plus	206	1,350	3,375	20,842
Qwen3-VL-Plus <sup>⚡</sup>	210	1,396	3,228	14,334
Gemini-2.5-Flash	225	4,388	3,124	3,343
Gemini-2.5-Flash <sup>⚡</sup>	206	2,025	3,043	3,022
Gemini-2.5-Pro <sup>⚡</sup>	198	2,576	2,501	2,451
GPT-4o-mini	270	71,274	230,971	-
GPT-4o	264	1,993	5,507	8,383
GPT-5 <sup>⚡</sup>	202	1,161	3,219	8,382

Table 8: Average prompt token consumption across different input modalities. <sup>⚡</sup> denotes inference with slow thinking. <sup>⊗</sup> indicates that videos are sampled at a maximum of 6 FPS due to the context window limitation.

decreases when slow thinking is enabled. These findings suggest that instruction-following robustness in sign language understanding is influenced by modalities, architectural and training choices, and may be further affected by the introduction of explicit reasoning.

## C.4 Case Studies

The qualitative cases in Figure 10 to Figure 14 provide intuitive insights into the sources of observed performance differences and further illustrate the challenges revealed by the CNSL-benchmark.

A primary observation concerns modality sensitivity. For the lexical concept *laptop* (from Figure 10 to Figure 12), models consistently succeed under textual descriptions but frequently fail under sign language video inputs, despite correctly recognizing the underlying typing motion. In contrast, image-based inputs sometimes enable correct predictions by exposing complementary cues (e.g., typing combined with an “open” gesture). This contrast indicates that sign language understanding is strongly affected not only by the presence of visual information, but also by how spatial and temporal cues are presented across modalities. Beyond modality effects, the cases suggest that sign language comprehension often requires integrating perceptual recognition with implicit semantic associations. As shown in Figure 13, Gemini-2.5-Flash may correctly identify individual sign components (e.g., a nose-related gesture corresponding to “smell”) while failing to connect them to the intended meaning in context. Such failures reflect the fact that sign language is not a simple composition of isolated gestures, but relies on culturally and linguistically grounded conventions that are not always explicitly observable. At the same time,



Figure 10: Text-based sign language understanding for the lexical concept “laptop”. Under textual descriptions, models successfully associate the described typing action with the intended concept, demonstrating robust performance when sign language information is abstracted into natural language.

we observe emerging sign-specific capabilities in stronger models. In Figure 14, Gemini-2.5-Pro correctly maps a visual sign to the Chinese character “neng” (meaning “can” or “ability”), indicating partial acquisition of sign knowledge.

Conclusively, these examples emphasize that the difficulties faced by current MLLMs stem primarily from the intrinsic linguistic and multimodal complexity of sign language, rather than from isolated perceptual errors alone.

		笔记本 (Laptop)		A.笔记本电脑	A. Laptop
		(一) 双手五指弯曲, 指尖朝下, 交替点动几下, 如敲击计算机键盘状。 (二) 双手横伸, 掌心相贴, 然后右手做向上打开的动作。 (1) Bend the fingers of both hands, with the fingertips pointing down, and alternately tap a few times as if striking the keys of a computer keyboard. (2) Stretch both hands out horizontally, with the palms touching each other, and then make an upward opening motion with the right hand.		B. 勇士	B. Warrior
				C. 因为 (由于)	C. Because/Due to
				D. 传讯	D. Transmit
	Qwen3-VL-2B	首先得分分析图示内容。图示有两部分, ..... 选项D是“传讯”, 传讯在口语中通常有特定手势。比如, 传讯可能..... C是“因为”, “因为”在口语中..... <b>选项A是“笔记本电脑”, 但图示里没有出现电脑, 所以排除。</b> 选项B是“勇, 勇士”的手势可能有特定的姿势..... 选项D是“传讯”, ..... 再仔细看图, (一)是两组手部动作, (二)是另一组。可能(二)的动作是手指向某个方向, 比如传递信息。再检查其他选项, A是笔记本电脑, 但图示没有出现电脑, 排除。B是勇士, 勇士的字可能需要特定手势, 比如“勇”字的手势, 但图示中可能不是。所以综合考虑, <b>正确选项应该是D。</b>			
	Qwen3-VL-8B	首先看选项, C是“因为(由于)”。口语中“因为”通常用手势表示原因, <b>比如一只手表示原因, 另一只手表示结果, 或者有特定的手势动作。</b> 图中可能有表示原因的手势动作, 比如(一)可能是手部动作。(二)是另一部分, 结合常见口语, “因为”是常用的口语表达。而其他选项, A笔记本电脑、B勇士、D传讯, 这些对应的手势应该不太符合图中的动作。比如“传讯”可能有传递信息的手势, 但图中动作更像表示原因。 <b>所以正确选项是C。</b>			
	GLM-4.1V	首先得回忆中国口语中各手势的含义。先看选项C“因为(由于)”, 需要判断手势是否匹配。不过可能需要更准确的分析, 不过根据常见口语, 这里可能对“因为”? <b>不对, 再想。</b> 或者看图示动作, 不过可能我之记忆有误差? 或者再仔细想, 可能正确选项是C? <b>不对, 再检查。</b> 哦, 可能正确答案是C? 不, 再想。或者看选项中的“因为(由于)”的手势动作? <b>不对, 可能我之前错了。</b> 重新想: 中国口语中, “因为”的手势动作? 或者看选项C是“因为(由于)”, 而图示如果是这个手势, 那选C? 或者可能我之记忆错了? 现在要确定唯一正确选项。假设正确选项是C, 所以输出C? <b>不对, 再想。</b> 或者看其他选项, A是笔记本电脑, 手势不太像; B勇士, 手势也不像; D传讯, 也不对。 <b>那剩下的C可能正确。所以最终选C?</b>			
	Gemini-2.5-Flash	..... Part (一) shows two hands, fingers slightly curled, like they're grabbing something or maybe vibrating over a wavy line. The wavy line is a common indicator of repetitive action or something happening. .... There's an arrow indicating this descending motion. Laptop signs usually have clear handshapes resembling an open device. .... "Because/Due To" is the strongest contender. Part (二), with the dominant hand descending ..... Putting it all together, the two parts align perfectly with the structure of "because" (因为) in CSL: 1. Part (一) sets the context, presenting the situation/reason. 2. Part (二) shows the dominant hand signifying the reason placing into the palm of the other hand representing the result. <b>Considering this, option C, "因为 (because/due to), looks like the clear winner.</b>			
	Gemini-2.5-Pro	..... First, I see two distinct movements. Movement (一) involves a typing motion, palms down, fingers bent, hands moving up and down slightly. That's clearly representing typing, almost universally. Next, Movement (二) depicts opening and closing, a hand moving down to cover another, shaped to mimic a laptop closing. It's very clearly the laptop itself. Combining those, it has to be a "laptop". I know how CSL, and indeed many sign languages, often use iconic gestures ..... D. 传讯 (chuānxùn) - To send a message/dispatch: Could involve information, but the specificity of the gestures to a laptop are too precise. Message uses index fingers. Based on that, I'm confident in "A". <b>Therefore, the final output needs to be "A".</b>			

Figure 11: Image-based sign language understanding for “laptop”. With illustrative images, some models correctly infer the target meaning by jointly recognizing multiple visual cues (e.g., “type” and “open” gestures), highlighting the role of complementary spatial information in sign language interpretation.

		笔记本 (Laptop)		A.笔记本电脑	A. Laptop
		(一) 双手五指弯曲, 指尖朝下, 交替点动几下, 如敲击计算机键盘状。 (二) 双手横伸, 掌心相贴, 然后右手做向上打开的动作。 (1) Bend the fingers of both hands, with the fingertips pointing down, and alternately tap a few times as if striking the keys of a computer keyboard. (2) Stretch both hands out horizontally, with the palms touching each other, and then make an upward opening motion with the right hand.		B. 勇士	B. Warrior
				C. 因为 (由于)	C. Because/Due to
				D. 传讯	D. Transmit
	Qwen3-VL-2B	视频中, 人物先是站立, 然后开始做手部动作。..... 选项里有“因为(由于)”这样的词。..... 首先分析手部动作: 在视频中, 手部动作可能是“因为”或者“由于”的手势。比如, 口语中“因为(由于)”的手势是通过手指的动作来表达, 比如用双手的特定手势来表示原因。现在看选项, C选项是“因为(由于)”, 所以需要确定是否符合。根据常见口语, 比如“因为”或“由于”的手势动作, 通常会有类似双手交叉或者特定的手势。 <b>根据视频中的动作, 人物先做出双手分开, 然后交叉, 这可能是“因为”的手势。所以, 正确选项应该是C。</b>			
	Qwen3-VL-8B	..... 选项中有A. 笔记本电脑, B. 勇士, C. 因为(由于), D. 传讯。..... <b>“因为”可能用手势表示为“手在胸前做某种动作, 可能类似表示原因”的动作。</b> 观察视频中的动作: 从6秒开始, 双手动作, 可能是在做“因为”的手势。比如, 手在胸前做交叉或指向等动作, 这可能对应“因为”。再看选项, C是“因为(由于)”。 <b>A笔记本电脑通常用手势表示电子设备, 可能不匹配:</b> B勇士可能有特定手势, 比如手臂伸展; D传讯可能涉及传递信息的动作, 但根据视频中的动作描述(6-11秒的手势), 更可能的是“因为”。需要确认标准口语中“因为”的手势。 <b>常见的“因为”手势是: 一只手放在胸前, 另一只手做类似“原因”的手势, 比如手掌朝上, 手指弯曲, 然后做某种表示因果的动作。</b> 视频中第6-10秒的动作可能对应这个。 <b>所以正确选项应该是C。</b>			
	GLM-4.1V	用户现在需要解决的问题是判断视频中手部动作对应的选项。首先得回忆常见口语动作的含义。 <b>选项C“因为(由于)”的手势动作特征, 或者看动作的对应。</b> 但可能需要分析动作流程, 不过通常这种选择题如果有正确答案的话, 需要确定。 <b>假设正确选项是C? 不对, 可能需要再想。或者可能我之前有记忆, C选项“因为(由于)”的手势动作表现。所以最终选C?</b>			
	Gemini-2.5-Flash	..... The signer starts with relaxed hands at their sides. At the two-second mark, the hands move up, palms facing each other. Then, at three seconds, things get interesting. <b>The left hand forms a flat palm, facing upwards, and the right hand forms a flat palm facing downwards. And, importantly, the right hand taps the left hand once. After that, the hands return to the relaxed starting position. .... Tapping one flat hand on top of another, or placing one flat hand on another, usually signifies "because," "reason," or "due to." .... A: 笔记本电脑 (Laptop computer) No, that's usually shown by forming a "laptop" shape with the hands or using a keyboard motion. Doesn't match. B: 勇士 (Warrior) Definitely not. That would involve gestures of strength, fighting, or holding a weapon. Doesn't fit. C: 因为 (由于) Bingo! This gesture aligns perfectly with the hand movement in the video. The foundation/reason concept is right on target. D: 传讯 (Transmit a message / Message) Doesn't seem right. That would involve a "sending" gesture or something similar to writing. <b>Therefore, my conclusion is that the sign in the video most likely represents **C: 因为 (Due to).</b></b>			
	Gemini-2.5-Pro	..... First, I need to really look at the video itself. <b>Hands down to start, then they come up. Ah, palms down, fingers extended, and together. The right hand goes on the back of the left hand, forming a sort of base, right? It's held in front of the body, a bit lower than chest height. This really screams "foundation" to me. .... A. Laptop Computer: Nope. The sign for "laptop" just doesn't involve this handshape or movement. It's much more dynamic, mimicking opening and closing a laptop. That's out. B. Warrior/Brave: No way. "Brave" usually has a fist gesture, near the chest or shoulder, signifying strength. Totally different. <b>C. Because/Due To: Wait a second... That's it! That's exactly the sign.</b> The left hand is the foundation, and the right hand is placed on it, showing the basis for something. Yes, that's spot on. D. Transmit a Message/Summon: Nope. The sign for "transmit" usually involves a movement away from the face or head, showing the sending of information. This isn't that at all. <b>So it's gotta be "Because/Due To", option C.</b></b>			

Figure 12: Video-based sign language understanding failure for “laptop”. Despite correctly recognizing the typing motion in a continuous sign language video, models fail to recover the intended meaning, indicating that temporal visual recognition alone is insufficient for reliable sign language understanding.

		<p><b>天气 / 空气 (Function)</b></p> <p>(一) 一手食指直立, 在头一侧上方转动一圈。        (二) 一手打手指字母“Q”的指式, 指尖朝内, 置于鼻孔处。</p> <p>(1) <i>Raise the index finger of one hand and extend it vertically; then perform a circular motion above one side of the head.</i>        (2) <i>Form the handshape corresponding to the manual alphabet letter "Q" with the fingertips oriented inward and positioned at the level of the nostrils.</i></p>	<p>A. 服务器  <b>B. 天气 (空气)</b>        C. 宗派        D. 芒果</p>	<p>A. Server  <b>B. Weather/Air</b>        C. Sect        D. Mango</p>
	<p>用户现在需要分析手语动作表达的意思。首先看视频中的动作：<b>第2到7秒之间, 手语动作包括拍手食指指向上(可能表示“点”或“开始”), 然后食指靠近嘴边(可能表示“小心”或“注意”), 接着挥手动作。这些动作组合起来, 手语中常见的是表达“天气(空气)”吗? 不, 再想, 可能对应的是“服务器”(A)? 但更常见的是“天气(空气)”可能对应B? 不对, 再看手语的常见含义。不过用户需要明确, 根据中国手语中“天气(空气)”这个说法是否对应。可能手语动作中, 先指向上(比如点), 然后靠近嘴边(可能表示“注意”), 挥手可能表示“空气”相关, 但不太对。再想, 可能“天气”是B选项, 但需要确认。不, 正确的分析应该是: 手语动作中, 指向上(可能表示“点”), 然后食指靠近嘴边(可能表示“注意”), 挥手可能表示“空气”? 不对。可能正确的选项是B.天气(空气)? 不, 再仔细查看选项, B是天气(空气), 可能对应手语动作中表示天气或空气的含义。或者可能, 手语动作是“天气”对应的, 比如用手指动作表示天气。但可能正确的是B。不过, 我需要确认。其实, 常见手语中, “天气”可能对应的是某个动作, 比如“点”(手指向上)表示“天气”, 然后其他动作。<b>但具体来说, 可能视频中的动作是“天气(空气)”, 即B选项。</b></b></p>			
	<p>用户现在需要分析手语动作对应的选项。首先看视频里的动作: 从画面看, 手语动作可能是“天气(空气)”。<b>因为手语中“天气”相关的动作可能涉及手势指向天空或表示空气。选项B是天气(空气), 所以选B。</b></p>			
	<p>用户现在需要解决手语动作对应的含义问题。首先得回忆手语中各选项对应的动作, B选项“天气(空气)”相关的手语动作, 可能需要分析动作手势, 但这里可能更直接的是识别常见手语。<b>通常天气或空气类手语动作可能与手的位置、动作表达空气相关, 而其他选项如服务器、宗派、芒果的手语动作不太符合。所以判断为B。</b></p>			
	<p>Analysis of the CSL Sign: Okay, so the user wants me to figure out what this sign means. <b>Let's break it down: right hand up, fist closed, then down to the nose, with the index finger and thumb pinching.</b> Classic "smell" or "nose" in Chinese Sign Language (CSL). Easy enough. Now, the options. Server? Nope. That's a "serve" and "machine" kind of deal. Sect? "Religion" and "group," definitely not. Mango? They're showing me cutting and peeling imaginary fruit, totally off. <b>But Weather/Air? Hmm... "Air" usually has an open hand, like a breeze, or moving around the face. The pinching's not quite right for "air" on its own. However... what if it's about the smell of the air? Pollution, maybe? That's a possibility, but let's see.</b> Looking at some signs for Weather and Air in CSL, I see this gesture isn't a direct sign for those. Okay, let's keep digging. Searching for signs for each concept shows that it's just the sign for nose, or smell, or the act of smelling. If I am being specific to the movement the options aren't right, yet it does appear that there's a connection. Mango is definitely out. Server and Sect are no go. Now for "Weather/Air". This is the only one I can even try and connect the sign to. In terms of CSL, is there any association between it and the nose or the act of smelling? <b>"Smelling the air"</b> could be linked. Given that this is a <b>multiple</b> choice quiz, and the sign means smell and the best option to pick would be "Weather/Air", let's review the sign again. This sign means nose or smell, and in the CSL. <b>So out of the options given, picking "Weather/Air" is the closest and only reasonable interpretation in the given context.</b> The sign is definitively "smell," and that strongly suggests "空气 (air)" when thinking about air quality or the sensory experience of the air (does it smell bad?). <b>Even if it's not the primary sign, this is the best fit. I'm going with "B. 天气 (空气)".</b></p>			
	<p>Alright, here's what I'm thinking. I'm being asked to identify the sign language displayed in the video. Time to break it down systematically, as I always do. First, I need to really analyze the video. Okay, so the woman starts with a fist near her shoulder, thumb up. Immediately, that shouts "sky" or "heaven" (天) to me. Classic. Then, she brings that closed hand to her nose. That's a clear "air" or "breath" (气) gesture, potentially even "air" (空气) on its own depending on the context. Now for the crucial part: putting it all together. It's almost certain that combining those two signs gives us "weather" (天气). Next, I need to evaluate the options provided, starting with A, "服务器 (fú wù qì) - Server." Nope, the signs for "serve" or "server" (as in the machine) are totally different; probably involve flat hands or motions related to computing. Doesn't fit. B, "天气 (空气) - Weather (Air)"... Bingo! The combination of "sky" and "air" directly translates to "weather" (天气), or even "air" (空气) on its own. It's a perfect match. C, "宗派 (zōng pài) - Sect/Denomination." That's an abstract concept. I'd expect signs related to "group" or "belief" for this, nothing close to what's in the video. Wrong. And finally, D, "芒果 (máng guǒ) - Mango." No way. The sign for mango would involve hand shapes depicting the fruit or the action of eating it. Completely off base. <b>Therefore, the video demonstrably displays the sign for "weather" or "air". The answer, without a doubt, is B.</b></p>			

Figure 13: Partial semantic recognition without correct contextual integration. Models correctly identify individual sign components (e.g., a nose-related gesture), but fail to integrate them into the intended semantic concept, reflecting the implicit and convention-driven nature of sign language semantics.

	<p><b>功能 (Function)</b></p> <p>(一) 左手食、中指与右手食指先搭成“工”字形，然后右手食指在左手旁书空“力”字，仿“功”字形。</p> <p>(二) 一手直立，掌心向外，然后食、中、无名、小指弯曲一下。</p> <p>(1) First, the index and middle fingers of the left hand and the index finger of the right hand form the shape of the character “gōng” (工). Then, the index finger of the right hand writes the character “lì” (力) in the air beside the left hand, imitating the shape of the character “gōng” (功).</p> <p>(2) Hold one hand upright with the palm facing outward, and then bend the index, middle, ring and little fingers once.</p>	<p>A. 蓝</p> <p>B. 撒拉族</p> <p>C. 短语 (短句)</p> <p>D. 功能</p>	<p>A. Blue</p> <p>B. The Salar nationality</p> <p>C. phrase</p> <p>D. Function</p>
	<p>用户现在需要解决手语图示的意思。首先看选项，选项A是“蓝”，B是“撒拉族”，C是“短语 (短句)”，D是“功能”。手语图示 (一) 和 (二) 的描述需要理解。手语通常是表达特定含义，可能涉及字词或短语。但看选项中的C是“短语 (短句)”，而手语里的“短语”可能是表达句子结构。不过需要分析图示。通常，手语图示可能用于表示某种短句或词汇。不过需要看选项。可能这个手语图示是表示“短语”或者“短句”的意思。因为C选项是“短语 (短句)”。其他选项：A“蓝”是颜色，手语可能有颜色词，但图示看起来是手势动作。B“撒拉族”是民族，一般用拼音或文字，手语可能涉及民族，但图示不是这个。D“功能”是功能，比如工具功能，但图示是手语手势。再仔细看，手语图示可能表达的是“短语”或者“短句”。因为选项C是“短语 (短句)”，而正确答案可能是C。因为短语 (短句) 在手语中常用手语表达。所以选C。</p>		
	<p>用户现在需要解决手语图示的问题。图 (一) 是左手食指和右手食指的动作，可能表示某种手势；图 (二) 是手掌张开，手指弯曲，有箭头指示动作。然后看选项：A蓝，B撒拉族，C短语 (短句)，D功能。首先，手语中“短语”或“短句”可能对应特定手势。比如，可能用手指的动作表示分句，或者重复手势。另外，“功能”是更抽象的概念，可能手势不同。撒拉族是民族，蓝是颜色，通常颜色用手语中“蓝”对应的手势是蓝色的表示，但需要具体手势。再想，图 (一) 可能是手势的组合，图 (二) 是手掌动作。可能“短语 (短句)”的手语是表示短句的结构，比如用手指划动表示短句。或者查常见手语词汇，“短语”可能对应某种手势。另外，选项C是“短语 (短句)”，D是“功能”。需要回忆中国手语中“短语”或“短句”的表达。比如，短语可能用手势表示分隔，短句可能重复或特定动作。而“功能”可能手势不同。可能图 (一) 是手势示意，图 (二) 是手掌动作，组合起来表示“短语”或“短句”。例如，在手语中，“短语”可能用手指在空中划动表示，而“短句”可能类似。或者“短语”对应特定手势。再想，可能这个图示是“短语 (短句)”的手语表达，因为其他选项不太符合。比如A蓝：颜色蓝的手语通常是用蓝色的表示，比如“蓝”字的手势，而图中动作不是；B撒拉族：民族的手语可能用特定手势，比如“撒拉族”可能用指代的动作，但图中不太像；D功能：功能的手语可能用“功能”这个词的表达，可能不同。所以综合判断，选项C更可能正确。</p>		
	<p>用户现在需要解决手语图示对应的意思，选项是A到D。首先得回忆中国手语中各个手语的表达。看选项C是短语 (短句)，需要判断图示是否符合。不过可能需要结合手语知识，通常这类手势如果是表示短语 (短句) 的手势？或者再仔细想，可能正确选项是C？不对，再检查。或者可能我之前记错了？不对，重新分析。用户提供的图示，如果是表示“短语 (短句)”的手语，那选C。但需要确定每个选项对应的手语。假设正确答案是C，那输出C？不对，可能我之前有误解？或者再想，可能正确选项是C？哦，不对，再仔细分析问题。选项C是短语 (短句)，所以判断后选C？不对，可能我记错了，再想。或者可能正确选项是C？最终确定选C？不对，再检查。哦，可能正确答案是C？所以最终输出C？</p>		
	<p>Alright, let's break this down. The task is to decipher these Chinese Sign Language gestures. I'm seeing two distinct signs in the image, labeled (一) and (二). Analyzing (一), the first gesture, it looks like a "scissors" shape with the index and middle fingers extended on one hand. The other hand's index finger is moving towards those "scissors." There's even a "力" symbol, hinting at force or emphasis during the movement. My initial thought is that this could represent a kind of "cutting" or articulation, maybe relating to separating something. This could potentially represent the concept of "word" or "phrase" in CSL. Now for (二). This one shows a hand with fingers extended, but then they curl down into a fist, or at least a relaxed position with the fingertips pointing down. The downward curling indicates something is shortening. This is a very clear representation of "short" in CSL. Now, putting it all together. Considering the options, let's eliminate the outliers first. "蓝" (Blue) is definitely out, as it would likely involve gestures related to the sky or eyes. "撒拉族" (Salar ethnic group) is unlikely, as it's too specific. And "功能" (Function) has nothing to do with shortening or words. That leaves us with "短语 (短句)" - "Phrase (Short Sentence)". This is where it gets interesting. "短" (duǎn, short) definitely maps to gesture (二). "语" (yǔ, language/speech/phrase) is the key. Thinking about it, the movement in gesture (一) with the "scissors" and the index finger could be interpreted as "articulating" or "separating" words. The "力" symbol would then reinforce the concept of force or emphasis. It represents words coming out in some way. I'm pretty confident that the image depicts the signs for "短" and "语" - or "句", in the context of sentences. Let me double-check my internal knowledge of CSL. Remembering CSL resources, "短" (short) is often a downward movement of the fingers, which matches (二) perfectly. "语" (word/phrase) often involves a tapping or movement between the fingers. Gesture (一) represents a form of articulation by using the index finger on the "scissors" like the mouth or the sound of words coming out. I'm pretty confident in the final answer. Therefore, I'm going with C, "短语 (短句)".</p>		
	<p>Alright, here's what I'm thinking. I'm being asked to decipher a Chinese Sign Language gesture. Okay, let's break it down methodically. First, I need to see what's going on in the image. (一) shows a hand using the "scissors" handshape, with another hand tracing the character "力" (lì), meaning "power" or "function". I recognize that. Then, in (二), the handshape changes to three fingers extended, and then they bend downwards. That signals the character "能" (néng), meaning "can" or "ability". Now, time to combine those. "力" and "能" together immediately bring to mind "功能" (gōng néng), which translates to "function" or "capability". So, I'll go through the options, just to be thorough. A. 蓝 (lán - blue): Nope, that's not related. The sign for "blue" would look entirely different. B. 撒拉族 (Sālázú - Salar people): No, the Salar sign wouldn't use those handshapes or movements. C. 短语 (短句) (duǎnyǔ (duǎnyǔ) - phrase/short sentence): Nope. The sign for "phrase" or "sentence" is a different gesture altogether. D. 功能 (gōngnéng - function): Bingo! That's the combination of "力" and "能". Therefore, the answer is D.</p>		

Figure 14: Emerging sign-specific symbolic understanding in advanced MLLMs. A stronger model successfully maps a visual sign to the Chinese character “neng” (meaning “can” or “ability”), suggesting partial acquisition of sign language knowledge, while still falling short of comprehensive understanding.