

Testing Video MLLM Sensitivity to Hand-Localized Visual Detail in Chinese Sign Recognition: A Matched Visual Intervention Study

Bowen Ye¹ Dengfeng Yao^{1,2,3} Siyuan Jing²
Peng Jin² Hongzhe Liu¹ Cheng Xu¹

¹ Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, China

³ Lab of Computational Linguistics, School of Humanities, Tsinghua University, Beijing, China

² Sichuan Province Key Laboratory of Philosophy and Social Science for Language Intelligence in Special Education, Leshan Normal University, Leshan 614000, China

Corresponding author: Dengfeng Yao (tjtdengfeng@bnu.edu.cn)

Abstract

Recent benchmarks show that multimodal large language models (MLLMs) can answer questions about sign-language video, but accuracy alone does not reveal which visual cues support those answers. We test behavioral sensitivity to hand-localized visual detail using 150 front-view isolated signs from NationalCSL-DP. For each item, a hand-target condition (T) blurs tracked hand regions and a same-arm control (C) applies the same Gaussian operator, nominal σ , and translated mask geometry to nearby arm regions; the underlying regional content is not information-matched. The same four-choice question is used across conditions and four video MLLMs. All 1,800 final Primary responses are uniquely scorable, with 1,200 T/C responses entering the primary contrast. Mean accuracy is 38.17% on original videos (O), 33.50% on T, and 36.67% on C. The pre-specified C–T contrast is +3.17 percentage points (95% bootstrap CI $[-0.17, 6.50]$; item-cluster sign-flip $p = 0.08172$). The four model-level point estimates are descriptively positive, but two models are floor-limited. The evidence is therefore directional rather than confirmatory. The contribution is a matched-control diagnostic that complements aggregate benchmark accuracy while making its residual anatomical-content mismatch explicit.

1 Introduction

Signed languages are full natural languages whose lexical and grammatical contrasts are expressed through coordinated manual and non-manual signals. For isolated signs, fine hand appearance—including handshape and local orientation cues—can be central, while a video model can also exploit easier regularities in the signer, body, background, or overall motion. A correct multiple-choice answer therefore does not by itself show that the

model was sensitive to the visual detail that linguistically distinguishes the sign.

CNSL-bench evaluates 21 open- and closed-source MLLMs on Chinese National Sign Language and documents large capability gaps across modalities and articulatory categories (Zhao et al., 2026). Recent sign-language MLLM work also spans linguistically informed modeling and small-VLM translation (Inan et al., 2025; Chowdhury and Sanyal, 2026). These results establish what systems can answer, but leave a narrower question open: are video-MLLM decisions selectively sensitive to fine visual detail localized on the hands?

We test that question by perturbing the input rather than changing the task. Each sign is rendered in three conditions: an unmodified standardized video (O), a hand-target condition (T), and a same-arm control (C). T and C use the same Gaussian operator and nominal severity, but at different anatomical locations (Figure 1). If hand-localized detail matters more than a nearby local perturbation, accuracy should be lower under T than C. This is a behavioral diagnostic, not a claim to recover an internal causal representation.

The study makes three contributions. First, it turns isolated-sign recognition into a cue-sensitivity diagnostic rather than another model-ranking exercise. Second, it introduces an anatomically constrained translated control with a pairwise-symmetric no-op rule. Third, it evaluates a frozen four-model panel with pre-specified paired and item-cluster inference. The Primary estimate is +3.17 points, but its interval includes zero and the sign-flip test remains above 0.05; we treat that uncertainty as part of the result rather than as something to optimize away.

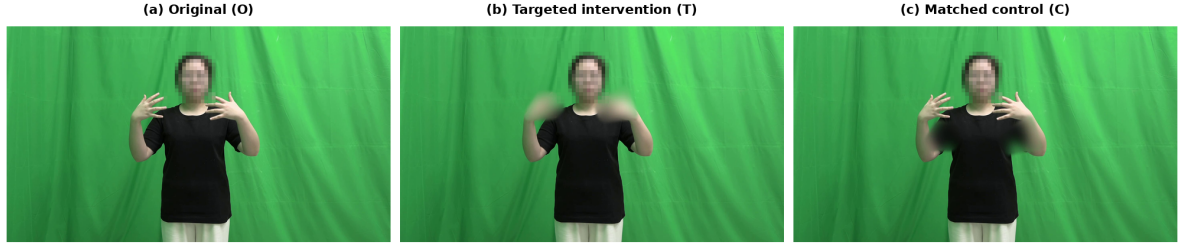


Figure 1: A frame from one of the 150 evaluation items. O is the standardized input; T applies Gaussian blur to tracked hand regions; C applies the same local operator to a translated copy of the target envelope along the same arm. The face is pixelated only in this publication figure; model inputs were unchanged.

2 Related Work

Signed languages should not be reduced to generic gesture streams (Yin et al., 2021). Sign-language MLLM studies expose modeling, data, and evaluation challenges, including signer-dependence effects (Inan et al., 2025; Chowdhury and Sanyal, 2026; Artiaga et al., 2025). Broader video-MLLM diagnostics reveal temporal and motion-specific weaknesses through MVBench, TempCompass, Video-MME, and MotionBench (Li et al., 2024; Liu et al., 2024; Fu et al., 2025; Hong et al., 2025). Together, they motivate diagnostics of what visual-temporal information supports a decision, rather than accuracy alone.

Our closest benchmark context is CNSL-bench (Zhao et al., 2026). We do not extend its leaderboard; we hold the item and question fixed and locally alter a hypothesized evidence source. Related to contrast-set evaluation (Gardner et al., 2020), our change is label-preserving and anatomically paired: the scientific quantity is the *difference between two local degradations*, not the drop caused by blur alone.

3 Matched Visual Intervention

Data, sampling, and task. NationalCSL-DP contains 6,707 Chinese National Sign Language signs and 134,140 dual-view videos from ten signers (Jin et al., 2025). We use Participant 01’s front-view stream, making this a controlled within-signer diagnostic rather than a signer-generalization test. The pre-outcome protocol removed explicit variant-suffix entries and three media-only calibration signs, leaving 6,026 eligible entries. A frozen PCG64 draw selected 180 targets; the first 150 were FORMAL and the remaining 30 RESERVE. All source videos passed the media gate, so no reserve item was used. Each four-choice item pairs the target with three random distractors from the same

Design factor	Control and residual scope
Local corruption	Same Gaussian operator and nominal σ ; regional content is not information-matched.
Mask geometry	Translate the target envelope; anatomical texture and motion may still differ by location.
Question/options	Fix prompt, candidate set, and option order within each item.
Invalid placement	Apply PAIRWISE-SYMMETRIC-NOOP to both T and C for that hand-frame.
Outcome selection	Freeze items/media before scoring; no outcome-based replacement.

Table 1: What the C–T design controls, and what remains residual.

eligible pool under frozen identity and translation-uniqueness constraints; distractors were not semantic nearest neighbors. Candidate sets and order are identical across conditions and models, with correct positions A=38, B=38, C=37, D=37. Dataset-provided lexical labels are used as the answer key and were not independently re-annotated for this study. Exact construction seeds are in Appendix A.

Conditions and identification logic. O is the standardized source video. T applies Gaussian low-pass blur to tracked hand regions. C translates a copy of the same hand-shaped mask along the corresponding wrist–elbow–shoulder path and applies the same operator and nominal σ there (Figure 1). Thus T and C share the video, question, candidate set, mask geometry, operator, and nominal severity; they do *not* guarantee equal perceptual or model-relevant information removal because the underlying regional content differs. A positive $\text{Acc}(C) - \text{Acc}(T)$ therefore indicates greater behavioral sensitivity to the hand-localized placement under this matched-corruption design, not pure removal of linguistic information. O is contextual; O–T is not the primary estimand.

Geometry and degradation. Hand and arm locations are obtained with RTMW-L (Jiang et al., 2024) at 384×288 , followed by a five-frame temporal median. The 21 hand keypoints define a convex-hull core; the target envelope dilates it by $0.15 \times$ hand scale. Candidate controls translate a copy of this envelope along the same-arm path with $\gamma \in \{1.0, 1.1, \dots, 3.0\}$, selecting the smallest valid placement with hand-core overlap ≤ 0.05 , target/control IoU ≤ 0.10 , and area retention ≥ 0.98 .

Both T and C use Gaussian blur with $\sigma = 0.16 \times$ hand scale and boundary feathering with $\sigma = 0.04 \times$ hand scale. If no valid control exists for a hand-frame, that hand-frame is left unchanged in both T and C (PAIRWISE-SYMMETRIC-NOOP). Formal items require at least 98% pairwise-valid hand-frame coverage and no invalid interval longer than 0.15 s. In the final set, 55,172/55,248 hand-frames (99.8624%) are pairwise-valid; 76 use symmetric no-op. Minimum item coverage is 98.5119%, and the longest invalid interval is 0.12 s.

Pre-outcome media safeguards. Severity was selected only from fixed media calibration. At $\sigma = 0.16 \times$ hand scale, fine finger separation and local handshape edges were suppressed while coarse hand position and gross trajectory remained readable. A pre-outcome calibration-panel MAE sanity check was approximately 0.1845 for T versus 0.1698 for C; it was not an information-equivalence metric and motivates treating the corruption as nominally, not information-, matched. All 450 formal videos passed full decode/resolution/frame-count/frame-rate checks; O/T/C keep the same native rate (25 or 50 fps) within an item and 1920×1080 resolution. No outcome-based replacement occurred.

4 Experimental Setup and Inference

Prompt and execution. The frozen Chinese prompt presents four candidate glosses and requests one letter (A/B/C/D) with no explanation; exact wording and manifest hash are in Appendix A. Prompt bytes and option order are identical for O/T/C and across providers. The execution manifest is blinded to the correct letter and target gloss; the scoring key is separately sealed. Condition order is balanced across all six O/T/C permutations (25 items each), and each provider follows a pre-generated item order with an item’s three conditions adjacent.

The completed Primary panel contains Qwen, GLM, Doubao, and ERNIE; each receives all 150 items in all three conditions. All 1,800 final Primary responses are uniquely scorable (malformed count: 0), with 1,200 T/C responses entering the estimand. Each scheduled call seeks one scientific response; retries are infrastructure-only, never quality resampling. Frozen decoding/request details are in Appendix A. Provider video ingestion is partly unobservable, so absolute accuracies are descriptive rather than a cross-provider leaderboard.

Gemini and MiniCPM-V were predeclared secondary systems but were not closed before the scoring key was opened, so later secondary execution is recorded as a sequence deviation. Gemini remained partial after debinding (20 valid responses; the rest infrastructure-failed or not run), and MiniCPM-V never reached formal video execution; neither altered Primary requests, panel membership, or statistics.

Inference. The primary estimand is $\Delta = \text{Acc}(C) - \text{Acc}(T)$. Per-model paired comparisons use exact McNemar tests with Holm adjustment across four tests. For each item, d_i is the four-model mean of $\text{correct}(C) - \text{correct}(T)$. The confirmatory panel-level test flips the sign of d_i at the item-cluster level (100,000 Monte Carlo flips; seed 20260819031). Under the null, these item-level signed contrasts are treated as sign-symmetric around zero for this fixed four-model panel; inference therefore targets the sampled items under this panel, not a population of video MLLMs. A 10,000-resample item-cluster percentile bootstrap (seed 20260819032) gives the 95% interval. A pre-specified floor guard flags models with $\text{Acc}(O) \leq 0.30$; flagged models remain equally weighted in the fixed panel.

5 Results

All 150 items enter the panel-level analysis, and all 600 model-by-item T/C pairs are complete. Mean accuracy is 38.17% on O, 33.50% on T, and 36.67% on C. Relative to O, T lowers mean accuracy by 4.67 points and C by 1.50. The Primary C–T estimate is +3.17 points, with a bootstrap 95% interval of $[-0.17, +6.50]$ and sign-flip $p = 0.08172$; the interval includes zero and the confirmatory test does not cross 0.05.

Descriptively, C–T is positive for each model (Table 2). C-correct/T-wrong versus T-correct/C-wrong counts are 16/7 for Qwen, 5/3 for GLM,

Model	O	T	C	C-T
Qwen	46.67	37.33	43.33	+6.00
GLM	28.00	27.33	28.67	+1.33
Doubao	48.00	44.00	47.33	+3.33
ERNIE	30.00	25.33	27.33	+2.00

Table 2: Primary results. O/T/C are accuracies (%); C-T is in percentage points. GLM and ERNIE trigger the pre-specified floor guard.

26/21 for Doubao, and 8/5 for ERNIE; unadjusted exact McNemar p values are 0.09314, 0.72656, 0.56006, and 0.58105, and none survives Holm adjustment. The interval and panel-level sign-flip result are the primary evidence; the four same-sign point estimates are descriptive context only. We therefore treat the result as directional, not confirmed.

6 Discussion

The C-T comparison is stronger than a clean-versus-blurred ablation because it holds the local Gaussian operator, nominal σ , mask construction, question, and options fixed while moving the perturbation anatomically. It does not, however, make the two regions information-equivalent. Hands contain different spatial detail and motion structure from nearby forearm regions, and the pre-outcome media sanity check itself shows a modest $T > C$ pixel-difference asymmetry. A generic sensitivity to the amount of local detail removed can therefore contribute to C-T. The estimand is best read as selective behavioral sensitivity to the hand-localized placement under a nominally matched corruption, not as a pure measure of linguistic evidence use.

The +3.17-point estimate is uncertain: its 95% interval crosses zero and two models are floor-limited near the 25% chance level. The fixed-panel test and interval are more informative than sign agreement across four point estimates. Retaining the floor-limited models is deliberate: the Primary estimand is the unweighted mean over the pre-specified panel, and removing them after observing outcomes would change that estimand.

The operator also has a narrower construct than “hand evidence” broadly. Gaussian low-pass blur suppresses fine appearance detail while largely preserving coarse hand position and gross trajectory, so the contrast does not isolate handshake from all other fine manual appearance cues. Provider-side resizing and temporal preprocessing are partly unobservable and may attenuate or transform the

effective perturbation before inference. A stronger follow-up would pre-specify a larger model panel, evaluate multiple signers, and use interventions that separately target fine handshake/orientation cues, coarse trajectory/location, and complementary non-manual evidence.

7 Conclusion

Across 150 isolated Chinese signs and a fixed four-model video-MLLM panel, the pre-specified C-T contrast is +3.17 percentage points (95% CI $[-0.17, +6.50]$; sign-flip $p = 0.08172$). The interval includes zero, so the result is directional rather than confirmatory; the four positive model-level point estimates are descriptive rather than separate confirmatory evidence. Matched visual interventions can complement benchmark accuracy by testing sensitivity to hypothesized cue locations while making the control’s residual mismatches explicit.

8 Limitations

The study covers one signer, 150 isolated signs, one dataset, and a four-choice task, so it does not establish signer generalization, continuous signing, discourse, or open-ended translation. Dataset labels were not independently re-annotated, and fixed pre-outcome visual/geometric QC was not an independent ground-truth tracking study. The control does not match regional information content; provider resizing/frame sampling is partly unobservable; and no full-sample native-pixel σ distribution is asserted. Two Primary models are floor-limited. The fixed four-model panel does not support broad generalization across video MLLMs. The incomplete post-deblinding secondary branches do not enter Primary inference.

9 Ethical Considerations

We reuse NationalCSL-DP and recruit no new participants; the dataset paper reports informed consent and ethics-board approval for public release (Jin et al., 2025). The representative face in Figure 1 is pixelated only for publication; model inputs are unchanged. We frame errors as system limitations and encourage Deaf and sign-language user involvement in future accessibility studies.

Acknowledgments

This work was supported by the General Project of the Beijing Social Science Fund (“Research on the Intangible Cultural Heritage Attributes and

Living Transmission of Local Sign Languages: A Case Study of Beijing Sign Language,” Grant No. 25BJ03128); the 2024 Key Project of the Sichuan Province Key Laboratory of Philosophy and Social Science for Language Intelligence in Special Education (“Processing Specifications for Phonological and Morphological Annotation of Sign Language: Corpus-Based Annotation Practice,” Grant No. YYZN-2024-2); and the General Project of the National Language and Writing Promotion Base (Beijing Union University) (“Research on Automatic Multichannel Phonological Annotation Methods for Sign Language Corpora for Information Processing,” Grant No. YW202504).

Generative AI tools were used to assist with code development, numerical cross-checking, and language and editorial refinement during manuscript preparation. All substantive research decisions, experimental design, analyses, interpretation of results, citation verification, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the work.

References

- Keren Artiaga, Sabyasachi Kamila, Haithem Afli, Conor Lynch, and Mohammed Hasanuzzaman. 2025. [Re-thinking sign language translation: The impact of signer dependence on model evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18379–18391. Association for Computational Linguistics.
- Anal Roy Chowdhury and Debarshi Kumar Sanyal. 2026. [Can small vision–language models perform sign language translation?](#) In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 32150–32166. Association for Computational Linguistics.
- 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. 2025. [Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323. Association for Computational Linguistics.
- Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. [MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8450–8460.
- Mert Inan, Anthony Sicilia, and Malihe Alikhani. 2025. [SignAlignLM: Integrating multimodal sign language processing into large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3691–3706. Association for Computational Linguistics.
- Tao Jiang, Xinchun Xie, and Yining Li. 2024. [RTMW: Real-time multi-person 2D and 3D whole-body pose estimation](#). *Preprint*, arXiv:2407.08634.
- 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*, 12:660.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. [MVBench: A comprehensive multi-modal video understanding benchmark](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. [TempCompass: Do video LLMs really understand videos?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Rui Zhao, Xuwen Zhong, Xiaoyun Zheng, Jinsong Su, and Yidong Chen. 2026. [CNSL-bench: Benchmarking the sign language understanding capabilities of MLLMs on Chinese national sign language](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40869–40890. Association for Computational Linguistics.

A Execution Details

Models and frozen request settings. Qwen: qwen3.7-plus-2026-05-26, Base64 video_url, exposed fps=10, temperature=0, max tokens=16, thinking disabled, stream=false. GLM: glm-5v-turbo, Base64 video_url, provider-native video preprocessing, temperature=0, max tokens=16, thinking disabled, stream=false. Doubao: doubao-seed-2-0-pro-260215, frozen Base64-video route, provider-native preprocessing, temperature=0, max tokens=16, thinking disabled, stream=false. ERNIE: ernie-4.5-turbo-v1, MP4/mp4v Base64 data URL, exposed fps=5, use_audio=false, temperature=0, max tokens=16, stream=false; no explicit thinking parameter was recovered. Parameters not explicitly set (including top-p and a decoding seed) were left to provider defaults.

Task construction. The source vocabulary had 6,707 rows; frozen eligibility left 6,026. Target sampling used NumPy Generator/PCG64 seed 20260818 (180 without replacement; 1–150 FOR-MAL, 151–180 RESERVE), correct-position balancing seed 2026081801, and distractor seed 2026081802. Distractors came from the same eligible pool with target exclusion and normalized-English-translation uniqueness constraints, not semantic nearest-neighbor selection. All 180 candidate glosses mapped uniquely to Participant_01-front/CSY/front and passed the media gate, so no replacement was used.

Exact model-visible prompt. The 150-record UTF-8 prompt manifest has SHA-256 75aa014aea247476aad81f97097d79b1e9aa2ba823cd589996a8c58c099b1d10. Its fixed instruction is:

你将看到一段国家通用手语的孤立词视频。请仅根据视频内容，从以下四个候选词中选择最符合该手语表达含义的一项。只输出一个大写字母A、B、C或D，不要解释，不要输出任何其他内容。

Pinyin transliteration: Ni jiang kan dao yi duan Guojia Tongyong Shouyu de guli ci shipin. Qing jin genju shipin neirong, cong yixia si ge houxuan ci zhong xuanze zui fuhe gai shouyu biaoda hanyi de yi xiang. Zhi shuchu yi ge daxie zimu A, B, C huo D, bu yao jieshi, bu yao shuchu renhe qita neirong.

English translation: You will see a video of an isolated word in Chinese National Common Sign Language. Based only on the video content, choose from the following four candidate words the one that best matches the meaning expressed by the sign. Output only one uppercase letter A, B, C, or D. Do not explain and do not output anything else.

The four candidate glosses then follow as A–D. For a given item, prompt bytes and A/B/C/D order are identical across O/T/C and all Primary models.

Retries and parsing. Each scheduled call has at most three attempts: attempt 2 after 5 s and attempt 3 after an additional 20 s, only for transient transport/network/timeout/408/425/429/5xx failures. Retries keep the same model, video, prompt, and provider parameters. HTTP-200 empty/refusal/safety/malformed outputs are scientific outputs and are not retried for quality; permanent non-transient 4xx errors are infrastructure failures. Returned text is NFKC-normalized with whitespace removed; an exact A/B/C/D or unique leading answer is accepted, while conflicting/non-unique outputs are malformed scientific outputs. Final Primary malformed count is zero.