FingerCap: Fine-grained Finger-level Hand Motion Captioning

Xin Shen[1,2], Rui Zhu[1], Lei Shen[3], Zhuojie Wu[2], Xinyu Wang[2], Kaihao Zhang[4], Tianqing Zhu[5], Shuchen Wu[1], Chenxi Miao[1], Weikang Li[1*], Deguo Xia[1], Jizhou Huang[1], and Xin Yu[2,6⋆]

[1] Baidu Inc, Beijing, China
[2] The University of Queensland, Brisbane, Australia
xin.shen@uq.edu.au
[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[4] Australian National University, Canberra, Australia
[5] City University of Macau, Macao, China
[6] Adelaide University, Adelaide, Australia

Abstract. Understanding human hand motion requires reasoning that goes beyond coarse action semantics to capture detailed finger articulation and interaction dynamics. However, existing Video-MLLMs primarily capture global action intent and often fail to represent fine-grained finger-level movements and contact semantics. In this work, we propose Fine-grained Finger-level Hand Motion Captioning (**FingerCap**), which aims to generate textual descriptions that capture detailed finger-level semantics of hand actions. To support this task, we curate **FingerCap-40K**, a large-scale corpus of 40K paired hand-motion videos and captions spanning two complementary sources: concise instruction-style finger motions and diverse, naturalistic hand–object interactions. Furthermore, we design **HandJudge**, an LLM-based rubric that measures finger-level correctness and motion completeness for effective evaluation. To establish a strong yet lightweight baseline under the FingerCap task, we introduce **FiGOP** (Finger Group-of-Pictures), a module tailored for Video-MLLM settings with sparse RGB sampling. FiGOP augments keyframes with intermediate 2D hand keypoints, enabling complementary motion cues without increasing the density of RGB frames. Experiments on **FingerCap-40K** show that strong open- and closed-source Video-MLLMs still struggle with finger-level reasoning, while our FiGOP-enhanced model yields consistent gains under both **HandJudge** and human studies. The dataset and benchmark are available at FingerCap.

Keywords: Finger-level Hand Motion · Motion Captioning Dataset · Video-Language Models

1 Introduction

Human hand motion is critical for both physical manipulation [20, 21, 36, 48, 76] and nonverbal communication [2, 9, 26, 35]. From grasping tools and typing [19,

⋆ Corresponding author.

Fig. 1: FingerCap aims to generate textual descriptions that capture detailed finger-level semantics of hand actions. Examples from **FingerCap-40K**: top, concise instruction-style clips with explicit targets for finger articulation; bottom, hand–object interactions showing coordinated finger dynamics during manipulation.

40, 46, 50] to signing and gesturing [3, 65, 83], the hands convey rich semantic and functional information. However, most existing research focuses on coarse hand-level actions [21, 76] or global gestures [9, 54], overlooking the subtle yet crucial contributions of individual fingers. Fine-grained finger articulation plays a critical role in modeling dexterity, precision, and intent, where subtle configuration changes may result in different gesture semantics or influence task success [7,21,37,40,50]. To bridge this gap, we introduce **Fine-grained Finger-level Hand Motion Captioning (FingerCap)**, a new task that generates detailed textual descriptions of how individual fingers move and coordinate during hand actions. Unlike conventional motion captioning [12,21,38,39,63] or gesture recognition [2, 26, 35, 54], FingerCap requires models to capture and describe fine-grained finger articulation, temporal evolution, and inter-finger coordination, whether in gesture communication or object manipulation.

To support the FingerCap task, we curate **FingerCap-40K**, a large-scale dataset of 40K video–caption pairs. As illustrated in Figure 1, the dataset spans two complementary domains: gesture instruction and hand–object interaction (HOI). The gesture domain is built from sign language datasets across four regions [2, 26, 35, 54], providing linguistically structured examples of fine finger articulations that are reviewed and refined by sign language experts. Captions in these samples offer diverse hand configurations and compositional gestures with explicit semantic meaning, serving as high-quality supervision for finger-level understanding [10, 62]. In contrast, the HOI domain captures physically grounded behaviors such as grasping, twisting, pinching, and transferring objects, collected from large-scale multi-view datasets [21,76] and out-of-distribution benchmarks [25, 48]. This domain complements the gesture data by introducing natural, unconstrained finger coordination in manipulation tasks. By integrating linguistically precise gestures with physically diverse interactions, FingerCap-40K provides both semantic richness and physical realism, forming a comprehensive foundation for modeling fine-grained finger motion captions.

Under typical sparse RGB sampling strategies adopted by current Video-MLLMs [4, 8, 13, 27, 42, 67, 68, 71, 77, 81], fine-grained finger articulations are

usually underrepresented. To establish a simple yet effective baseline in this setting, we introduce **FiGOP** (Finger Group-of-Pictures), a lightweight module that associates each sparsely sampled RGB keyframe with a temporally adjacent sequence of 2D hand keypoints [75]. A temporal graph-based encoder (ST-GCN) [69, 74] aggregates the sequence into a compact motion representation, which is fused with visual tokens in the multimodal projector. This structured pose augmentation can be seamlessly integrated into existing Video-MLLM pipelines as a task-driven baseline for finger-level reasoning. Compared with increasing RGB sampling density, FiGOP preserves high-frequency motion cues while maintaining computational efficiency and scalability to long sequences.

To enable reliable evaluation of finger-level motion captions, we introduce **HandJudge**, a compact LLM-based assessment framework [23, 29, 34]. Conventional captioning metrics [6, 43, 51, 70] are insufficient for FingerCap, as improvements in lexical overlap do not necessarily reflect correct finger articulation, motion direction, or contact semantics. Without task-specific evaluation criteria, models may produce fluent yet physically inaccurate descriptions, masking the true difficulty of fine-grained motion reasoning. HandJudge evaluates generated captions along four dimensions: (1) Finger and Hand Identification (FHI), assessing the accurate recognition of both the acting hand and specific fingers; (2) Motion and Trajectory (MT), measuring the correctness of the movement type and its directional path; (3) Contact and Interaction (CI), evaluating the fidelity of physical interactions while strictly penalizing hallucinated or physically implausible contacts; and (4) Completeness of Motion Sequence (CMS), verifying whether the description captures all semantically essential phases defined in ground-truth annotations.

Through comprehensive experiments on both open- and closed-source Video-MLLMs [13, 27, 67, 71, 77], as well as task-adapted fine-tuned models, we observe that current models perform poorly on FingerCap. They frequently miss or misrepresent finger articulations, and in many cases, the generated captions are vague or even hallucinated. These results expose a fundamental gap in fine-grained hand motion understanding, emphasizing the necessity of dedicated benchmarks and specialized modeling strategies.

In summary, our contributions are fourfold:

- We define **FingerCap**, a new task for understanding and describing detailed finger articulation and coordination.
- We curate **FingerCap-40K**, a 40K-sample dataset combining linguistically precise gesture data and physically grounded hand–object interactions.
- We introduce **FiGOP**, a lightweight and memory-efficient module that binds sparse sampled RGB keyframes with dense 2D hand keypoints to capture temporal details.
- We propose **HandJudge**, an LLM-based evaluation framework for interpretable and multi-dimensional assessment of fine-grained motion semantics.

2 Related Work

2.1 Hand Motion Datasets

Research on hand motion understanding has led to the development of diverse datasets across three major domains: hand–object interaction, gesture recognition, and sign language recognition (ISLR) [53, 55, 56, 58–61]. Hand–object interaction datasets [5,20–22,47–49,76] have enabled progress in modeling physical manipulation and contact dynamics, supporting studies of grasping and coordination [19, 40, 50]. However, these datasets primarily focus on coarse action categories or hand trajectories, without offering detailed representations of finger articulation or semantic descriptions that capture intent. Gesture recognition and ISLR datasets, in contrast, focus on communicative and symbolic hand movements. While they vary in scale from dozens to thousands of gesture or gloss classes, they often remain limited to single-view RGB recordings, restricted viewpoints, or small vocabularies [9,26,44,45,52,64]. Although recent datasets have improved realism through larger signer diversity and higher resolution, they typically describe isolated gestures rather than continuous, fine-grained finger motion [2,33,35,54]. This limitation hinders the ability of these datasets to represent the subtle articulations and temporal coordination that underlie expressive and dexterous hand behavior [7]. Existing hand motion datasets have greatly advanced recognition and interaction understanding, but they largely neglect continuous finger-level dynamics and lack natural language grounding. To address this gap, our FingerCap-40K dataset provides large-scale paired video–text data covering both structured gestures and natural hand–object interactions, enabling a new research direction in fine-grained finger motion captioning.

2.2 Human Motion Understanding

Traditional motion understanding methods are predominantly built on 3D skeletal representations, where actions are modeled as sequences of articulated joints [24, 30, 41, 72, 73, 80]. These datasets and models are typically annotated at the body or hand level, without explicit supervision for individual fingers. As a result, they can capture coarse motion patterns but are fundamentally unable to learn how specific fingers articulate, coordinate, or interact with objects. Even hand-centric datasets and models [21,48,76] follow this design and describe hand pose at a single rigid unit, rather than providing finger-level trajectories. To enrich geometric modeling with visual cues, recent works combine RGB video and pose sequences [12, 28, 38, 39, 63]. However, the underlying annotations still operate at the hand level, and pose streams are usually sampled at low temporal resolution. Consequently, high-frequency finger movements and subtle transitions between finger configurations are either missing from the data or smoothed out during aggregation, preventing these models from developing true finger-level hand motion understanding.

Video Multimodal Large Language Models (Video-MLLMs) [13, 27, 57, 67, 71, 77, 81, 82] introduce powerful language reasoning on top of visual features

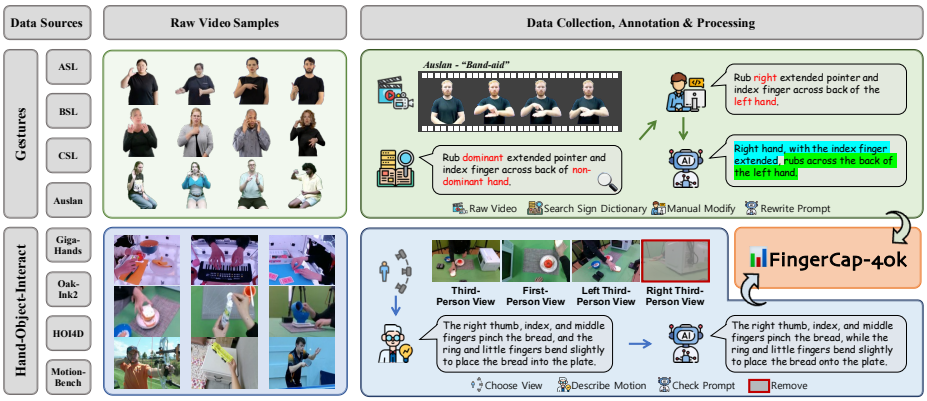

Fig. 2: Data collection, annotation and processing pipeline for gesture and hand–object interaction data in **FingerCap-40K**. Gesture videos are collected from multilingual sign language datasets, where raw dictionary-style motion descriptions are manually corrected and refined using an LLM to produce finger-level captions. Hand–object interaction videos are sampled from multi-view manipulation datasets, in which the clearest view is selected, followed by human-written and LLM-refined finger–object interaction descriptions.

and show emerging ability to describe hand actions from large-scale web data. However, they inherit the same limitations of their training corpora: sparse frame sampling and predominantly hand-level supervision. They can often infer the overall intent of a gesture, but they rarely capture which fingers move, in what order, and how they make or break contact. To efficiently mitigate temporal sparsity and expose models to explicit finger supervision, we propose FiGOP, which enriches Video-MLLMs with fine-grained hand keypoints for finger-level understanding.

3 FingerCap-40K

Building reliable models for finger-level hand motion understanding requires data that jointly capture both *semantic intent* and *fine-grained kinematics*. However, existing datasets fall short in several aspects: (1) gesture corpora [9, 54] often contain high-level linguistic semantics but lack explicit descriptions of finger articulations; (2) hand–object datasets [20, 22] focus on manipulation dynamics yet rarely include natural language annotations that describe motion in finger-level detail; and (3) few resources [21, 76] provide temporally aligned motion–caption pairs with consistent left–right and contact annotations. These limitations hinder the study of fine-grained motion reasoning and the evaluation of multimodal models under precise physical supervision. To bridge this gap,

we curate **FingerCap-40K**[7], a large-scale video–caption dataset featuring 40K fine-grained hand motion–language pairs.

3.1 Data Sources

FingerCap-40K is constructed from two complementary domains: gesture instruction and hand–object interaction. The gesture subset is collected from four major sign language systems, including ASL [35], BSL [2], CSL [26] and Auslan [54]. These videos provide naturally aligned motion–text pairs with semantically precise finger articulations. All annotations are reviewed and refined with the assistance of sign language experts to ensure structural correctness and temporal consistency. The hand–object interaction subset captures physically grounded finger motions such as grasping, twisting and pinching, sampled from large-scale multi-view datasets such as GigaHands [21] and OakInk2 [76]. To evaluate generalization under distribution shifts, we further include out-of-distribution samples from HOI4D [48] and MotionBench [25]. Together, these two domains provide both semantic precision and kinematic diversity, forming a comprehensive foundation for fine-grained finger-level hand motion captioning.

3.2 Data Collection, Annotation and Processing

Based on the two domains introduced above, we construct FingerCap-40K through a unified pipeline (Figure 2) that ensures temporal alignment, semantic precision and natural language quality in all video–caption pairs.

Gesture data. For sign language videos, we first retrieve motion descriptions from official sign language dictionaries corresponding to each dataset. These raw descriptions are not directly suitable for captioning due to three issues: (1) hand references are often written as "dominant" and "non-dominant" rather than left and right, which creates spatial ambiguity; (2) some entries include non-motor content such as emotions or analogies unrelated to physical motion; and (3) the language is instructional and lacks the natural sentence structure expected in captions. To address this, we manually correct or remove ambiguous content, normalize left and right hand references, and then paraphrase the text using GPT-4.1 [1] while preserving the original motion semantics. The resulting captions are concise, fluent, and finger-level accurate.

Hand–object interaction (HOI) data. For HOI videos, we sample clips from multiple viewpoints, including first-person, third-person, left, right and top-down cameras, and only retain those in which all finger movements are clearly visible. Annotators then describe how each finger interacts with the object, focusing on articulation, contact and coordination between both hands when present. These descriptions are subsequently refined using GPT-4.1 [1] to ensure grammatical consistency and descriptive clarity.

To prevent factual errors, the LLM (GPT-4.1) is strictly restricted to textual polishing and is explicitly forbidden from altering semantics or adding new

[7] FingerCap-40K and all sources follow the **CC BY-NC-SA 4.0** license.

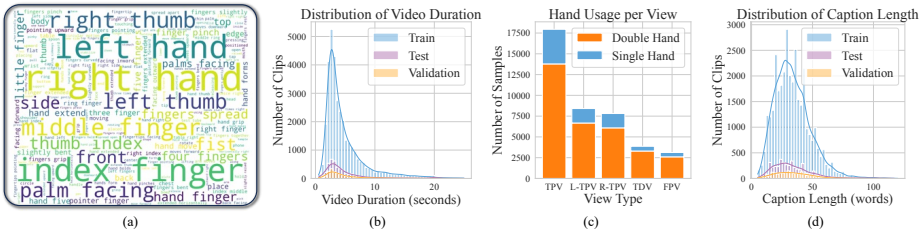

Fig. 3: Data distribution in **FingerCap-40K**. (a) the word cloud of finger- and hand-related terms in captions. (b) video duration. (c) single vs. double hand usage across viewpoints[8]. (d) caption length distribution.

information. Thus, the final captions are strictly grounded in human visual perception, completely avoiding LLM hallucinations.

3.3 Dataset Statistics

Table 1 presents the core statistics of FingerCap-40K, highlighting its scale and diversity. It contains 40K video with fine-grained finger-level hand motion captions, including 21K gesture clips and 19K hand–object interaction clips. In total, it comprises 6.17 million frames and 1.40 million caption words. The gesture subset has a vocabulary size of 3.4K, and the interaction subset has 4.8K unique words, indicating substantial linguistic diversity across communicative and manipulative domains. For training and evaluation, the dataset is split into training, validation, and test sets in a ratio of 8:1:1.

Table 1: Statistics of the **FingerCap-40K** dataset across gesture and hand–object interaction domains.

	Gesture	HOI
Data Source	ASL, CSL, Auslan	[21], [76]
Num.Videos	21,055	19,922
Num.Words	651,112	744,296
Num.Frames	1,922,471	4,251,389
Num.Vocs	3,427	4,882
Num.Views	3	5
Single-Hand	6,850	1,544
Both-Hand	14,205	18,378
OOD Set	BSL	[48], [25]
OOD Vocs	100	36
OOD Domain	sign language	sport, medical, ...

Figure 3 further analyzes the data distribution. The word cloud (a) shows frequent use of terms referring to individual fingers and hands, which reflects the dataset focus on fine-grained motion semantics. The video durations (b) are mostly between one and ten seconds, producing short but motion-rich clips. Caption lengths (c) exhibit a long-tailed distribution, indicating varied levels of descriptive complexity across samples. Hand usage across TPV, L-TPV, R-TPV, TDV and FPV (d) shows that bimanual actions occur more often than single-hand motions, offering a wide range of viewpoints that support modeling detailed finger movements under varied conditions.

4 FiGOP-augmented Video-MLLM

Current Video-MLLMs [4,13,15–18,67,68,71,77,81] typically adopt sparse RGB sampling to reduce computational cost when processing long videos. However, such sampling discards high-frequency motion cues, especially the rapid articulations and coordination of fingers. Inspired by the Group-of-Pictures (GOP) principle [32,78], we introduce **Finger Group-of-Pictures (FiGOP)**, a simple yet effective encoding mechanism that augments sparsely sampled RGB frames with dense hand pose streams. Unlike prior work that relies on optical flow or dense frame inputs [32, 78], FiGOP uses structurally organized 2D hand keypoints [75], which provide explicit motion trajectories without increasing pixel-level redundancy. An overview of this architecture is shown in Figure 4.

4.1 FiGOP Unit Construction

A video is divided into a sequence of FiGOP units. Each unit consists of:

$$\text{FiGOP}_t = (I_t,\, P_{t:t+K})\,, \tag{1}$$

where $I_t \in \mathbb{R}^{H\times W\times 3}$ is a sparsely sampled RGB keyframe, and $P_{t:t+K} = \{p_t, p_{t+1}, \ldots, p_{t+K-1}\}$ is a dense sequence of hand poses between the current and next keyframe. Each $p_i \in \mathbb{R}^{J\times C}$ represents J hand joints, with C-dimensional features (*e.g.*, (x, y) coordinates and confidence) [75]. This design preserves high-frequency motion while keeping RGB sampling unchanged.

4.2 Dual-Stream Encoding

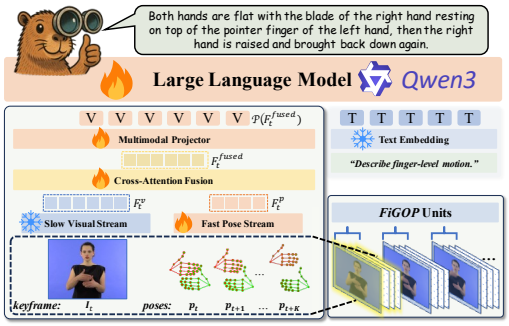

Fig. 4: Overview of the FiGOP-augmented Video-MLLM.

Each FiGOP unit is processed by two parallel streams:

- **Slow Visual Stream.** The RGB keyframe I_t is encoded by a pre-trained vision encoder [14,67,68] to obtain spatial tokens $F_t^v \in \mathbb{R}^{N\times D_v}$.
- **Fast Pose Stream.** The pose sequence $P_{t:t+K}$ is fed into the ST-GCN module [31, 58, 74] to model finger joint topology and local motion. A lightweight temporal Transformer [69] further aggregates cross-frame dependencies, producing $F_t^p \in \mathbb{R}^{K\times D_p}$.

Pose representations, unlike optical flow or RGB images, are structured and physically meaningful, offering computational efficiency and robustness to background and lighting variations [31, 58].

[8] TPV = third-person view, L-TPV = left third-person view, R-TPV = right third-person view, TDV = top-down view, FPV = first-person view.

4.3 Motion-Aware Projector Fusion

To inject high-frequency motion cues into the visual representation, we incorporate a motion-aware adapter [78] into the multimodal projector. Given visual tokens F_t^v and pose motion features F_t^p, we apply a cross-attention [69] fusion:

$$F_t^{fused} = \mathrm{CA}(F_t^v, F_t^p) + F_t^v, \tag{2}$$

$$\mathrm{CA}(F_t^v, F_t^p) = \mathrm{Attn}(F_t^v W_Q,\ F_t^p W_K,\ F_t^p W_V), \tag{3}$$

where W_Q, W_K and W_V denote learnable projection matrices. The fused tokens are then projected into the LLM (***Qwen3*** [68]) embedding space:

$$E_t^{LLM} = \mathcal{P}(F_t^{fused}). \tag{4}$$

Our design allows visual tokens to selectively retrieve motion information from pose features, improving finger-level temporal reasoning.

5 Experiments

5.1 Evaluation Metrics

Standard Caption Metrics. Following common practice in video captioning, we report BLEU-4 (**B-4**) [51], ROUGE-L (**R-L**) [43], METEOR (**M**) [6], and CIDEr (**C**) [70] (all values are multiplied by 100 for clearer comparison after using the *Jury* evaluation toolkit [11]). These metrics measure lexical overlap and fluency but do not effectively capture finger-level correctness, motion direction, or contact semantics, which are crucial for this task.

HandJudge (LLM-as-a-Judge). To address the limitations of standard metrics, we introduce HandJudge, an LLM-based evaluation framework [23, 29, 34] using GPT-4.1 [1]. For each generated caption, the LLM compares it with the ground truth and assigns a score from 0 to 5 across four expert-defined criteria: (1) finger and hand identification (**FHI**), (2) motion and trajectory accuracy (**MT**), (3) contact and interaction reasoning (**CI**), and (4) completeness of motion sequence (**CMS**). CMS evaluates whether the generated caption covers all semantically essential phases defined in the ground-truth annotation, rather than exhaustively describing every observable visual state in the video. The LLM also provides intermediate reasoning before scoring, offering interpretable and fine-grained assessment that better aligns with human judgment, as illustrated in Figure 5. More details are provided in the *Appendix*.

5.2 Implementation Details

Evaluation Protocol. We evaluate several open-source Video-MLLMs using LLaMA-Factory [79]. Closed-source systems are queried via APIs with unified

decoding settings. All models receive the same 2-fps-sampled RGB inputs, ensuring a fair comparison.

FiGOP-augmented Video-MLLM. We apply Qwen3-VL-8B [68] as the backbone, and use the efficient and precise DWPose [75] to extract 2D poses. Videos are sampled at 2 fps, and for each RGB keyframe, we attach a 2D hand-pose sequence in the following 8 frames, forming a FiGOP unit. RGB frames are encoded by the frozen vision tower [14], while poses are processed through a two-layer ST-GCN and temporal Transformer to generate pose motion embeddings.

Two-Stage Fine-tuning. We perform full supervised fine-tuning on FingerCap-40K. In Stage 1, we freeze the vision encoder and the LLM, and train the pose encoder and the projector for one epoch. In Stage 2, we unfreeze the LLM and fine-tune both projector and LLM with a next-token-prediction loss for three epochs. The whole process is trained on eight NVIDIA A100 GPUs with a batch size of 1 and learning rates of 1e−4 (Stage 1) and 1e−5 (Stage 2). More details of prompts, experimental settings, and evaluation are provided in the *Appendix*.

5.3 Baseline Models

Closed-source Models. We evaluate three state-of-the-art proprietary multimodal systems: GPT-4o [27], GPT-4o-mini [27], and Gemini-2.5-Pro [13]. These models are capable of understanding videos and represent the frontier in commercial multimodal reasoning.

Open-source Models. We include several recent state-of-the-art open-source MLLMs: `LLaVA-NeXT-Video-7B` (-V-7B) [77], `InternVL3-8B` [81], `InternVL3.5-8B` [71], `Qwen2.5-VL-7B-Instruct` (-I) [67], and `Qwen3-VL-8B-Instruct` (-I) [68]. These models, trained on large-scale visual–textual alignment and temporal adaptation, serve as transparent and reproducible baselines for fine-grained motion understanding.

Fine-tuned Variants. To evaluate task adaptation, we fine-tune `Qwen3-VL-8B-Instruct` on the FingerCap-40K dataset under two configurations: (1) using a standard multimodal projector (MM Proj), and (2) using our proposed FiGOP-augmented projector (FiGOP), which incorporates structured pose representations through the spatial–temporal fusion.

5.4 Results

Zero-shot Evaluation: Closed-source vs. Open-source Models. We first compare the performance of closed-source and open-source models in both standard captioning metrics and HandJudge evaluation. As shown in Table 2 and Table 3, Gemini-2.5-Pro [13] outperforms all other models in both evaluation settings. In the standard metrics, it achieves the highest METEOR (34.28) and CIDEr (32.37) scores. Meanwhile, Gemini-2.5-Pro maintains a leading position in HandJudge with an overall score of 2.64, showing superior performance in FHI and CMS. In comparison, the open-source models, especially Qwen3-VL-8B-Instruct [68], show competitive performance but fall short of Gemini-2.5-Pro. Specifically, though Qwen3-VL-8B-Instruct achieves a decent average CIDEr

Table 2: Performance comparison of different models on standard captioning metrics in the **FingerCap-40K** test set.

Model	Gesture				HOI				Average			
	B-4	R-L	M	C	B-4	R-L	M	C	B-4	R-L	M	C
Closed-source Models												
GPT-4o	1.54	18.36	27.95	24.64	2.71	22.12	28.42	17.54	2.09	20.13	28.17	21.29
GPT-4o-mini	1.11	14.15	25.58	21.05	1.60	13.66	26.15	13.06	1.34	13.92	25.85	17.28
Gemini-2.5-Pro	3.31	27.13	35.13	29.69	3.30	24.46	33.33	35.05	3.31	25.87	34.28	32.37
Open-source Models												
LLaVA-NeXT-V-7B	1.41	19.32	27.02	13.85	2.78	21.73	28.23	11.63	2.05	20.46	27.59	12.80
InternVL3-8B	1.33	21.49	28.98	18.38	2.45	23.72	29.87	28.76	1.86	22.54	29.40	23.28
InternVL3.5-8B	0.93	19.97	27.91	16.45	2.33	22.82	30.15	17.26	1.59	21.31	28.97	16.83
Qwen2.5-VL-7B-I	1.24	20.67	26.15	23.84	2.58	23.10	27.77	34.91	1.87	21.81	26.92	29.06
Qwen3-VL-8B-I	1.98	19.85	30.30	23.10	2.79	19.27	30.01	35.92	2.36	19.58	30.16	29.14
Fine-tuned Models: Qwen3-VL-8B-Instruct												
+ MM Proj + SFT	7.84	31.17	32.87	89.89	13.42	36.15	36.36	126.73	10.48	33.52	34.51	107.28
+ **FiGOP** + SFT	**13.81**	**36.84**	**38.41**	**146.29**	**17.09**	**39.14**	**39.43**	**165.31**	**15.36**	**37.92**	**38.89**	**155.27**

score of 29.14, it still lags behind Gemini-2.5-Pro in HandJudge, with an overall score of 2.09 compared to 2.64 for Gemini-2.5-Pro. This clearly indicates that closed-source models [68,71,77] perform better at capturing both lexical fluency and fine-grained hand motion details in comparison to their open-source [13,27] counterparts.

Impact of Fine-tuning. In the standard captioning evaluation, the fine-tuned Qwen3-VL-8B-Instruct demonstrates significant improvements across all metrics. Meanwhile, fine-tuning with the standard multimodal projector (MM Projector) also results in a substantial increase in performance, and even surpasses Gemini-2.5-Pro. This indicates that task-specific fine-tuning can boost the model's ability to generate accurate captions. However, in the HandJudge evaluation, the fine-tuned model with MM Projector still lags behind Gemini-2.5-Pro. The model achieves an overall HandJudge score of 2.13 compared with 2.64 achieved by Gemini-2.5-Pro. This suggests that while fine-tuning has a significant impact on caption generation, the finger-level accuracy remains a challenge, highlighting the importance of incorporating rich and task-specific data for fine-grained motion understanding.

FiGOP-augmented Video-MLLM. The introduction of FiGOP significantly enhances the performance of fine-tuned model. The FiGOP-augmented model shows improvements in both HOI and gesture subset, surpassing Gemini-2.5-Pro in multiple HandJudge metrics. Specifically, FHI, MT, and CI scores reach 2.96, 2.50 and 2.45, respectively, and the overall HandJudge score increases to 2.74, bringing the performance of the open-source model on par with the closed-source model in several aspects. These results suggest that FiGOP better captures fine-grained finger motion, mitigating limitations observed in earlier fine-tuned

Table 3: Results of **HandJudge** evaluation across four dimensions: Finger and Hand Identification (FHI), Motion and Trajectory (MT), Contact and Interaction (CI), and Completeness of Motion Sequence (CMS).

Model	Gesture				HOI				Average				
	FHI	MT	CI	CMS	FHI	MT	CI	CMS	FHI	MT	CI	CMS	Overall
Closed-source Models													
GPT-4o	2.26	1.75	1.28	2.94	2.86	2.34	2.17	2.65	2.54	2.03	1.70	2.81	2.27
GPT-4o-mini	1.79	1.36	1.20	2.26	2.17	1.95	1.81	2.40	1.97	1.64	1.49	2.33	1.86
Gemini-2.5-Pro	2.91	2.13	1.97	**3.38**	2.53	2.45	2.40	**3.33**	2.73	2.28	2.17	**3.36**	2.64
Open-source Models													
LLaVA-NeXT-V-7B	1.31	1.13	0.89	1.97	1.59	1.21	1.14	1.87	1.44	1.17	1.01	1.92	1.39
InternVL3-8B	2.02	1.66	1.27	2.68	2.01	1.92	1.72	2.39	2.01	1.78	1.48	2.54	1.96
InternVL3.5-8B	2.05	1.59	1.24	2.86	2.24	1.99	1.84	2.73	2.14	1.78	1.52	2.80	2.06
Qwen2.5-VL-7B-I	1.69	1.30	0.78	1.86	1.92	1.67	1.56	2.09	1.80	1.48	1.15	1.96	1.60
Qwen3-VL-8B-I	2.14	1.47	1.16	2.67	2.64	1.93	1.92	2.92	2.38	1.69	1.52	2.79	2.09
Fine-tuned Models: Qwen3-VL-8B-Instruct													
+ MM Proj + SFT	2.40	1.75	1.63	2.50	2.45	1.94	1.99	2.43	2.42	1.84	1.80	2.47	2.13
+ **FiGOP** + SFT	**3.03**	**2.41**	**2.34**	3.11	**2.88**	**2.60**	**2.58**	3.01	**2.96**	**2.50**	**2.45**	3.06	**2.74**

models and narrowing the performance gap between open-source and closed-source models.

5.5 Discussion

Evaluation Reliability. We randomly sample 100 clips and collect captions from three systems: the zero-shot Qwen3-VL-8B [68], multimodal projector (MM Proj) fine-tuned model, and our FiGOP-augmented model. Each caption is scored by three independent judges: Qwen2.5-7B [66], GPT-4.1 [1], and human annotators, with human judgments gathered from 5 independent raters. The evaluations follow the four HandJudge criteria (FHI, MT, CI, CMS; 0–5). As shown in Table 4, the scores from GPT-4.1 closely match those of the human raters across all models, demonstrating the reliability of LLM-based evaluations. Our FiGOP-augmented model outperforms other systems and attains the highest average score. These results indicate that the FiGOP-augmented model yields the most structurally accurate, consistent, and interaction-aware descriptions. Further details on the evaluation setup and protocol are provided in the *Appendix*.

Table 4: Comparison of HandJudge scores from Qwen2.5, GPT-4.1, and human judges.

Models	Judges		
	Qwen2.5-7B	**GPT-4.1**	**Human**
Zero-shot	2.08	1.55	1.10
MM Proj	3.30	3.18	3.01
FiGOP	**3.52**	**3.61**	**3.74**

Impact of Pose Sequence Length. A critical hyperparameter in the FiGOP architecture is the temporal receptive field of the pose stream, defined by the

Table 5: Ablation on pose sequence length (T_p).

T_p	B-4	R-L	M	C
4 Frames	14.12	36.10	37.95	147.50
8 Frames	**15.36**	**37.92**	**38.89**	**155.27**
16 Frames	15.15	37.75	38.50	153.80

Table 6: Comparison of model performance on out-of-distribution subsets.

Model	B-4	R-L	M	C	HandJudge
Zero-shot	1.33	20.13	27.57	9.67	1.56
MM Proj	2.53	27.38	27.33	47.66	1.61
FiGOP	**3.11**	**30.47**	**29.15**	**52.21**	**2.04**

Table 7: Ablation on modality limitations and naive video scaling strategies. **R**: RGB visual stream; **P**: Pose stream.

Method	Modality	Resolution	FPS	B-4	R-L	M	C	HJ
Pose-only	P	-	8	6.10	21.20	24.10	62.10	1.25
Base MLLM	R	512^2	2	10.48	33.52	34.51	107.28	2.13
Medium Sampling	R	512^2	4	13.10	35.30	36.30	124.00	2.29
Dense Sampling	R	512^2	8	OOM	OOM	OOM	OOM	OOM
Low-Resolution	R	256^2	8	11.70	34.60	35.60	118.00	2.25
FiGOP (Ours)	**R + P**	$\mathbf{512^2}$	**2 (R) / 8 (P)**	**15.36**	**37.92**	**38.89**	**155.27**	**2.74**

sequence length (T_p) anchored to each sparse RGB keyframe. In the main paper, we set $T_p = 8$ (representing 8 pose frames per RGB keyframe). To empirically validate this choice, we conducted an ablation study across $T_p \in \{4, 8, 16\}$. As detailed in Table 5, constraining the pose window to $T_p = 4$ leads to a noticeable degradation across all metrics. This indicates that restricted temporal windows are insufficient to encapsulate the full kinematic trajectory of rapid finger articulations. Conversely, expanding the sequence to $T_p = 16$ yields no further empirical gains and marginally deteriorates performance. We attribute this to the introduction of temporal redundancy and an increased susceptibility to accumulated noise from the 2D pose estimator (*e.g.*, spatial jitter or occlusion artifacts). Such noise can overwhelm the lightweight temporal encoder, attenuating its focus on the most salient high-frequency dynamics immediately adjacent to the keyframe. Consequently, $T_p = 8$ establishes the optimal equilibrium between motion granularity, robustness to estimation noise, and modeling efficiency.

Generalization under Distribution Shifts. To further examine model robustness, we evaluate on two out-of-distribution (OOD) subsets from *BSL* (linguistic variation) and *HOI4D + MotionBench* (physical variation). We compare the zero-shot Qwen3-VL-8B [68], multimodal projector (MM Projector) fine-tuned model, and our FiGOP-augmented model. As shown in Table 6, all metrics drop notably compared to the in-distribution test set, reflecting the inherent difficulty of generalizing to unseen sign languages or novel manipulation domains. Nevertheless, our method consistently outperforms both zero-shot and standard fine-tuned baselines across all metrics, indicating that incorporating structured motion cues improves resilience to distribution shifts. However, performance gaps remain large, particularly in the HandJudge overall score, suggesting that fine-grained reasoning about unseen finger configurations and contact dynamics remains an open challenge.

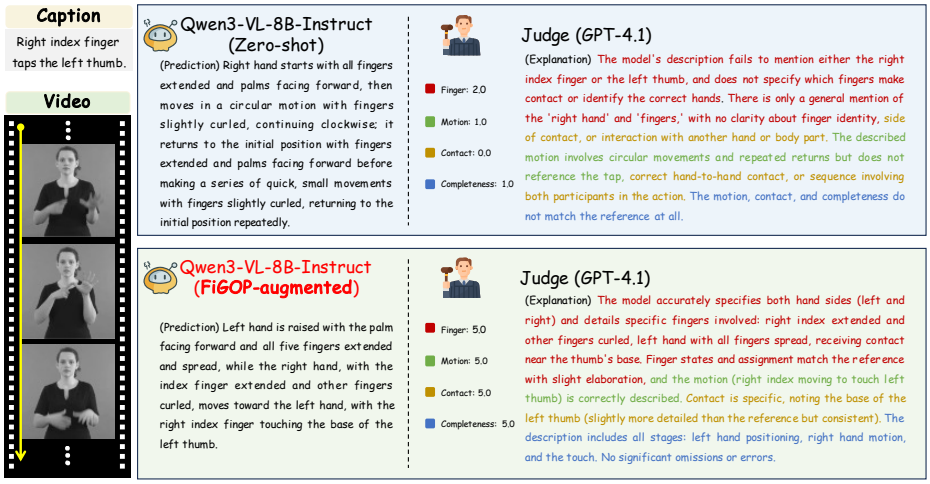

Fig. 5: Comparison of hand gesture descriptions between zero-shot and **FiGOP**-augmented models.

Ablation on Modality and Temporal Scaling. To justify the necessity of the FiGOP architecture, we evaluate its design against naive modality and temporal scaling alternatives in Table 7. First, relying solely on pose estimators without RGB context (*i.e.*, pose-only methods) fails completely, particularly in Hand-Object Interactions (HOI). Because identical 3D finger geometries can convey vastly different semantics depending on the unobserved objects, applying hard-coded geometric rules is practically infeasible. Second, naively scaling up the video temporal density by directly increasing the RGB sampling rate to 8 fps leads to severe Out-of-Memory (OOM) errors. Attempting to circumvent this by reducing the spatial resolution to 256×256 severely obscures fine-grained finger details, causing a noticeable performance regression. By elegantly embedding dense pose sequences (8 fps) between sparse, high-resolution RGB frames (2 fps), FiGOP breaks this spatial-temporal bottleneck, achieving optimal performance without triggering the memory explosion typical of dense Video-MLLMs.
Further Discussions. For more detailed analyses and additional ablation studies (such as the impact of different pose estimators), please refer to the *Appendix*.

5.6 Case Study

We compare the zero-shot model and our proposed FiGOP-augmented version of Qwen3-VL-8B-Instruct [68] on a hand gesture task involving the right index finger tapping the left thumb. The zero-shot model generates a vague description, failing to specify the fingers involved and the contact point. The GPT-4.1 [1] evaluation gives low scores, due to missing details such as finger identity and motion sequence. In contrast, the FiGOP-augmented model provides a detailed and

accurate description, specifying the left hand's position and the exact contact between the right index finger and left thumb. GPT-4.1 rates this highly, with perfect scores (5.0) in all dimensions. These results demonstrate the significant improvement in fine-grained motion description with the FiGOP augmentation. Further case studies can be found in the *Appendix*.

6 Conclusion

Understanding the subtleties of human hand motion requires a bridge between perception, language, and action. In this work, we take a step in that direction by introducing the **FingerCap** task, the first benchmark and framework for fine-grained finger-level motion description. Our **FingerCap-40K** dataset combines the linguistic precision of gesture with the physical realism of hand–object interaction, providing a foundation for studying how models interpret manual dexterity. We further propose the **FiGOP** module, which injects structured pose dynamics into video representations, enabling models to reason not only about visual appearance but also about temporal movement patterns at the finger level. Through evaluations using our **HandJudge** framework, we show that while existing Video-LLMs struggle with subtle motion understanding, incorporating explicit motion structure significantly narrows the gap. We hope this work inspires the community to move beyond coarse global actions and toward the rich, structured language of the human hand.

Acknowledgements

This research is funded in part by ARC-Discovery grant (DP220100800 to XY), ARC-DECRA grant (DE230100477 to XY) and Google Research Scholar Program. We also gratefully thank all the anonymous reviewers and ACs for their constructive comments.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 6, 9, 12, 14
2. Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J.S., Fox, N., Zisserman, A.: Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In: European conference on computer vision. pp. 35–53. Springer (2020) 1, 2, 4, 6
3. Arkushin, R.S., Moryossef, A., Fried, O.: Ham2pose: Animating sign language notation into pose sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21046–21056 (2023) 2
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 2, 8

5. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., et al.: Introducing hot3d: An egocentric dataset for 3d hand and object tracking. arXiv preprint arXiv:2406.09598 (2024) 4
6. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005) 3, 9
7. Bilge, Y.C., Cinbis, R.G., Ikizler-Cinbis, N.: Towards zero-shot sign language recognition. IEEE transactions on pattern analysis and machine intelligence **45**(1), 1217–1232 (2022) 2, 4
8. Cao, X., Xu, M., Yu, X., Yao, J., Ye, W., Huang, S., Zhang, M., Tsang, I., Ong, Y.S., Kwok, J.T., et al.: Analytical survey of learning with low-resource data: From analysis to investigation. ACM Computing Surveys **58**(6), 1–47 (2025) 2
9. Cao, X., Virupaksha, P., Jia, W., Lai, B., Ryan, F., Lee, S., Rehg, J.M.: Socialgesture: Delving into multi-person gesture understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19509–19519 (2025) 1, 2, 4, 5
10. Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., Johnson, T.: Signbank: Software to support web based dictionaries of sign language. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA) (2018), http://www.lrec-conf.org/proceedings/lrec2018/summaries/499.html 2
11. Cavusoglu, D., Sert, U., Sen, S., Altinuc, S.: Jury: A comprehensive evaluation toolkit (2023). https://doi.org/10.48550/arXiv.2310.02040 9
12. Chen, L.H., Lu, S., Zeng, A., Zhang, H., Wang, B., Zhang, R., Zhang, L.: Motionllm: Understanding human behaviors from human motions and videos. arXiv preprint arXiv:2405.20340 (2024) 2, 4
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 2, 3, 4, 8, 10, 11
14. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 8, 10
15. Du, X., Sun, H., Lu, M., Zhu, T., Yu, X.: Dreamcar: Leveraging car-specific prior for in-the-wild 3d car reconstruction. IEEE Robotics and Automation Letters **10**(2), 1840–1847 (2024) 8
16. Du, X., Wang, Y., Sun, H., Wu, Z., Sheng, H., Wang, S., Ying, J., Lu, M., Zhu, T., Zhan, K., et al.: 3drealcar: An in-the-wild rgb-d car dataset with 360-degree views. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26488–26498 (2025) 8
17. Du, X., Wang, Y., Yu, X.: Mvgs: Multi-view regulated gaussian splatting for novel view synthesis (2026), https://arxiv.org/abs/2410.02103 8
18. Du, X., Wang, Y., Zhan, K., Yu, X.: Mobile-gs: Real-time gaussian splatting for mobile devices. arXiv preprint arXiv:2603.11531 (2026) 8
19. Fan, Y., Yang, Q., Wang, K., Zhou, H., Li, Y., Feng, H., Ding, E., Wu, Y., Wang, J.: Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17550–17560 (2025) 1, 4

20. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12943–12954 (2023) 1, 4, 5
21. Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., Sridhar, S.: Gigahands: A massive annotated dataset of bimanual hand activities. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 17461–17474. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.01627, https://openaccess.thecvf.com/content/CVPR2025/html/Fu_GigaHands_A_Massive_Annotated_Dataset_of_Bimanual_Hand_Activities_CVPR_2025_paper.html 1, 2, 4, 5, 6, 7
22. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024) 4, 5
23. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, Y., Guo, J.: A survey on llm-as-a-judge. CoRR **abs/2411.15594** (2024). https://doi.org/10.48550/ARXIV.2411.15594, https://doi.org/10.48550/arXiv.2411.15594 3, 9
24. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022) 4
25. Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 8450–8460. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.00791, https://openaccess.thecvf.com/content/CVPR2025/html/Hong_MotionBench_Benchmarking_and_Improving_Fine-grained_Video_Motion_Understanding_for_Vision_CVPR_2025_paper.html 2, 6, 7
26. Huang, J., Zhou, W., Li, H., Li, W.: Attention-based 3d-cnns for large-vocabulary sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology **29**(9), 2822–2832 (2018) 1, 2, 4, 6
27. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 2, 3, 4, 10, 11
28. Hwang, M., Hejna, J., Sadigh, D., Bisk, Y.: Motif: Motion instruction fine-tuning. IEEE Robotics and Automation Letters (2025) 4
29. Jang, Y., Raajesh, H., Momeni, L., Varol, G., Zisserman, A.: Lost in translation, found in context: Sign language translation with contextual cues. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8742–8752 (2025) 3, 9
30. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023) 4
31. Jiao, P., Min, Y., Li, Y., Wang, X., Lei, L., Chen, X.: Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In: Pro-

ceedings of the IEEE/CVF international conference on computer vision. pp. 20676–20686 (2023) 8
32. Jin, Y., Sun, Z., Xu, K., Chen, L., Jiang, H., Huang, Q., Song, C., Liu, Y., Zhang, D., Song, Y., et al.: Video-lavit: unified video-language pre-training with decoupled visual-motional tokenization. In: Proceedings of the 41st International Conference on Machine Learning. pp. 22185–22209 (2024) 8
33. Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language. arXiv preprint arXiv:1812.01053 (2018) 4
34. Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al.: From generation to judgment: Opportunities and challenges of llm-as-a-judge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 2757–2791 (2025) 3, 9
35. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020. pp. 1448–1458. IEEE (2020). https://doi.org/10.1109/WACV45572.2020.9093512, https://doi.org/10.1109/WACV45572.2020.9093512 1, 2, 4, 6
36. Li, K., Li, P., Liu, T., Li, Y., Huang, S.: Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 6991–7003. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.00656, https://openaccess.thecvf.com/content/CVPR2025/html/Li_ManipTrans_Efficient_Dexterous_Bimanual_Manipulation_Transfer_via_Residual_Learning_CVPR_2025_paper.html 1
37. Li, K., Li, P., Liu, T., Li, Y., Huang, S.: Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6991–7003 (2025) 2
38. Li, L., Jia, S., Wang, J., An, Z., Li, J., Hwang, J.N., Belongie, S.: Chatmotion: A multimodal multi-agent for human motion analysis. arXiv preprint arXiv:2502.18180 (2025) 2, 4
39. Li, L., Jia, S., Wang, J., Jiang, Z., Zhou, F., Dai, J., Zhang, T., Wu, Z., Hwang, J.N.: Human motion instruction tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17582–17591 (2025) 2, 4
40. Li, M., Christen, S., Wan, C., Cai, Y., Liao, R., Sigal, L., Ma, S.: Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17416–17425 (2025) 1, 2, 4
41. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.T.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. arXiv preprint arXiv:2410.07093 (2024) 4
42. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Al-Onaizan, Y., Bansal, M., Chen, Y. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024. pp. 5971–5984. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.EMNLP-MAIN.342, https://doi.org/10.18653/v1/2024.emnlp-main.342 2
43. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 3, 9

44. Liu, D., Zhang, L., Wu, Y.: Ld-congr: A large rgb-d video dataset for long-distance continuous gesture recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3304–3312 (2022) 4
45. Liu, X., Shi, H., Chen, H., Yu, Z., Li, X., Zhao, G.: imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10631–10642 (2021) 4
46. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7037–7047 (2025) 1
47. Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., Yi, L.: Taco: Benchmarking generalizable bimanual tool-action-object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21740–21751 (2024) 4
48. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: HOI4D: A 4d egocentric dataset for category-level human-object interaction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 20981–20990. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.02034, https://doi.org/10.1109/CVPR52688.2022.02034 1, 2, 4, 6, 7
49. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12999–13008 (2023) 4
50. Pang, Y., Shao, R., Zhang, J., Tu, H., Liu, Y., Zhou, B., Zhang, H., Liu, Y.: Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12209–12219 (2025) 1, 2, 4
51. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002) 3, 9
52. Perera, A.G., Wei Law, Y., Chahl, J.: Uav-gesture: A dataset for uav control and gesture recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 4
53. Shen, X., Du, H., Sheng, H., Li, L., Zhang, K.: Auslanweb: A scalable web-based australian sign language communication system for deaf and hearing individuals. In: Proceedings of the ACM on Web Conference 2025. pp. 5212–5223 (2025) 4
54. Shen, X., Du, H., Sheng, H., Wang, S., Chen, H., Chen, H., Wu, Z., Du, X., Ying, J., Lu, R., Xu, Q., Yu, X.: Mm-wlauslan: Multi-view multi-modal word-level australian sign language recognition dataset. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/812c59ba55c03a68a10c25017bdb696e-Abstract-Datasets_and_Benchmarks_Track.html 2, 4, 5, 6
55. Shen, X., Du, H., Xu, M., Liu, M., Yu, X.: Cross-view isolated sign language recognition challenge: Design, results and future research. In: Companion Proceedings of the ACM on Web Conference 2025. pp. 2444–2447 (2025) 4

56. Shen, X., Ke, Y., Wang, X., Yu, X.: Banz-fs: Banzsl fingerspelling dataset. In: The Fourteenth International Conference on Learning Representations 4
57. Shen, X., Shen, L., Yuan, S., Du, H., Sun, H., Yu, X.: Diverse sign language translation. arXiv preprint arXiv:2410.19586 (2024) 4
58. Shen, X., Wang, X., Shen, L., Zhang, K., Yu, X.: Cross-view isolated sign language recognition via view synthesis and feature disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20647–20657 (2025) 4, 8
59. Shen, X., Yuan, S., Sheng, H., Du, H., Yu, X.: Auslan-daily: Australian sign language translation for daily communication and news. Advances in Neural Information Processing Systems **36**, 80455–80469 (2023) 4
60. Sheng, H., Shen, X., Du, H., Yu, X.: Mobile auslan: A multimodal dialogue-centered sign language learning system. Computer Vision and Image Understanding p. 104646 (2026) 4
61. Sheng, H., Shen, X., Du, H., Zhang, H., Huang, Z., Yu, X.: Ai empowered auslan learning for parents of deaf children and children of deaf adults. AI and Ethics **4**(4), 877–887 (2024) 4
62. Smith, R.T., Willoughby, L., Johnston, T.: Integrating Auslan resources into the language data commons of Australia. In: Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources. pp. 181–186. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.signlang-1.28 2
63. Song, G., Wang, G., Huang, Z., Lin, J., Zhe, X., Li, J., Wang, H.: Towards fine-grained human motion video captioning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 846–855 (2025) 2, 4
64. Starner, T., Forbes, S., So, M., Martin, D., Sridhar, R., Deshpande, G., Sepah, S., Shahryar, S., Bhardwaj, K., Kwok, T., et al.: Popsign asl v1. 0: An isolated american sign language dataset collected via smartphones. Advances in Neural Information Processing Systems **36**, 184–196 (2023) 4
65. Tang, S., He, J., Cheng, L., Wu, J., Guo, D., Hong, R.: Discrete to continuous: Generating smooth transition poses from sign language observations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3481–3491 (2025) 2
66. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/ 12
67. Team, Q.: Qwen2.5-vl (January 2025), https://qwenlm.github.io/blog/qwen2.5-vl/ 2, 3, 4, 8, 10
68. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388 2, 8, 9, 10, 11, 12, 13, 14
69. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 3, 8, 9
70. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015) 3, 9
71. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 2, 3, 4, 8, 10, 11

72. Wei, Y., Yuan, S., Chen, M., Shen, X., Wang, L., Shen, L., Yan, Z.: Mpp-net: multi-perspective perception network for dense video captioning. Neurocomputing **552**, 126523 (2023) 4
73. Wu, Y., Ji, W., Zheng, K., Wang, Z., Xu, D.: Mote: Learning motion-text diffusion model for multiple generation tasks. arXiv preprint arXiv:2411.19786 (2024) 4
74. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 7444–7452. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135 3, 8
75. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) 3, 8, 10
76. Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., Lu, C.: Oakink2 : A dataset of bimanual hands-object manipulation in complex task completion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 445–456. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00050, https://doi.org/10.1109/CVPR52733.2024.00050 1, 2, 4, 5, 6, 7
77. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), https://llava-vl.github.io/blog/2024-04-30-llava-next-video/ 2, 3, 4, 8, 10, 11
78. Zhao, Z., Huo, Y., Yue, T., Guo, L., Lu, H., Wang, B., Chen, W., Liu, J.: Efficient motion-aware video MLLM. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 24159–24168. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.02250, https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_Efficient_Motion-Aware_Video_MLLM_CVPR_2025_paper.html 8, 9
79. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, Bangkok, Thailand (2024), http://arxiv.org/abs/2403.13372 9
80. Zhu, B., Jiang, B., Wang, S., Tang, S., Chen, T., Luo, L., Zheng, Y., Chen, X.: Motiongpt3: Human motion as a second modality. arXiv preprint arXiv:2506.24086 (2025) 4
81. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 2, 4, 8, 10
82. Zhu, R., Shen, X., Wu, S., Miao, C., Yu, X., Li, Y., Li, W., Xia, D., Huang, J.: Video-msr: Benchmarking multi-hop spatial reasoning capabilities of mllms. arXiv preprint arXiv:2601.09430 (2026) 4
83. Zuo, R., Potamias, R.A., Ververas, E., Deng, J., Zafeiriou, S.: Signs as tokens: A retrieval-enhanced multilingual sign language generator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23806–23816 (2025) 2