Beyond Single-View Sufficiency: CVBench for Cross-View Human Understanding

Supplementary Material

A. Details of Construction

A.1. Details of Source Dataset

To ensure a comprehensive and rigorous evaluation of cross view human understanding, CVBench is constructed by sampling synchronized frames and video clips from a diverse pool of established public multiview datasets. This carefully curated selection provides a rich variety of scenarios encompassing indoor and outdoor settings, synthetic and physical data, and environments ranging from sparse to highly crowded scenes. Crucially, the selected clips feature four synchronized views and range from six to thirty seconds in length. This diversity is essential to guarantee that the benchmark comprehensively tests models on verifiable single view insufficiency, mandating the fusion of complementary and often conflicting visual cues. Below, we detail the specific source datasets utilized and their role in CVBench.

Ego-Exo4D [19]: This dataset captures skilled human activities from synchronized first person egocentric and third person exocentric perspectives. Only the exocentric videos from manually selected orthogonal views are utilized in this benchmark. It contains highly rich and diverse scenarios encompassing both indoor and outdoor environments. In CVBench, Ego-Exo4D is incredibly valuable for evaluating fine grained and coarse grained temporal tasks. The stark viewpoint disparity between egocentric and exocentric cameras mandates robust, spatially consistent geometric fusion alongside action recognition over time.

M3GYM [49]: A large scale multimodal, multiview, and multiple person pose dataset focused on fitness activity understanding in physical settings. It primarily features three main domains including normal, pilates, and yoga sessions. It contains 502 distinct actions in total across seven different lighting conditions, captured primarily via overhead camera angles. The complex indoor gym scenarios are characterized by overlapping individuals and severe self occlusions during exercise. This makes it an ideal source for constructing tasks like cross view human counting, repetition counting, and occlusion discrimination.

Wild-track [5]: Captured using multiple synchronized cameras, this high definition dataset is designed for dense, unscripted pedestrian detection. We utilize its crowded outdoor environments to construct coarse grained spatial tasks such as cross view trajectory summarization and identity association. These scenes directly challenge the tendency of models to double count or confuse visually similar individuals across different camera feeds, especially in expansive outdoor spaces.

MultiHuman [63]: This dataset focuses on the performance capture of multiple interacting characters using sparse multiview cameras. The close proximity physical interactions and complex intertwining of limbs provide the inherently ambiguous source material necessary to rigorously test cross view contact recognition and body part contact recognition.

Human-M3 [11]: A multiview and multimodal dataset developed for three dimensional human pose estimation, uniquely situated in outdoor scenes. It contributes unconstrained, natural environments that actively challenge the geometric grounding and identity persistence of a model when lighting conditions and backgrounds vary significantly across synchronized views.

MvMHAT [17]: Focusing on multiview human association and tracking, this dataset provides continuous video streams across multiple overlapping camera fields of view in outdoor environments. This continuous camera coverage is essential for our coarse grained temporal tasks, particularly cross view human counting and trajectory summarization, where a model must reconstruct a person's movement as they navigate through discontinuous visual fields.

MultiviewX [24]: A multiview detection dataset built around feature perspective transformation. It features synthesized outdoor scenarios with heavily dense pedestrian traffic. It also includes multiple identical digital avatars located in the same space but at different positions, which presents immense challenges for distinguishing individuals based solely on spatial positioning. Its highly crowded scenes are heavily utilized in CVBench to evaluate the robustness of models against single view bias during spatial reasoning tasks, actively penalizing naive detection aggregation. Due to the extreme density and the presence of identical digital humans in varying positions, Set of Mark prompting is applied for certain tasks to properly localize the specific individual in question.

EgoHumans [27]: An egocentric multiple human benchmark that captures individuals from synchronized first person and third person views. Only the third person video feeds are manually selected for this benchmark. It provides a rich and diverse set of both room and outdoor scenarios, supplying the spatio temporal data required for evaluating identity consistency and fine grained motion recognition as individuals move and interact within dynamic social environments.

A.2. View Selection Details

Each question and answer pair consists of four synchronized views. To ensure these views capture sufficient detail, provide diverse information, and remain challenging for the benchmark, we manually select four orthogonal views rather than relying on random sampling. For source datasets with dense camera coverage, we select the four most orthogonal viewpoints available. For instance, camera positions are chosen from the midpoints of four enclosing walls to capture maximum room information. For datasets with fewer views, such as MultiHuman, we select all available angles to capture as much diverse scene information as possible.

A.3. Distractor Details

For the human counting and action counting tasks, distractors are generated using single view observations, simple numerical additions of the given views, or the fusion of a partial subset of the provided views. For identity association, distractors are selected from individuals who are spatially closest to the target or whose appearance is most similar to the target. Specifically, for data sourced from Wild-track, the distractors involve the exact same individual appearing across all given images but situated in a spatially different position from the target. For contact recognition and occlusion discrimination, we utilize similar contact joints and surrounding objects, drawing results from single or subset views to act as distractors. For repetition counting, distractors are likewise selected from single views, subset views, and subsets of temporal frames to test whether the model fully comprehends the entire clip sequence. Finally, for motion recognition, action order, and trajectory summarization, distractors are derived from visually similar motions or by adding, removing, or disturbing the correct motion sequence. This ensures the models can successfully synthesize temporal information from multiple inputs and correctly distinguish disturbed temporal events.

A.4. Details of SoM Usage

Set of Mark prompting is exclusively applied to dense scenarios where standard text descriptions would introduce unavoidable ambiguity, such as scenes with more than twenty pedestrians in Wild-track and MultiviewX, which account for approximately 14 percent of the benchmark. The remainder of the benchmark does not use visual prompting, ensuring the overall results reflect core reasoning capabilities rather than sensitivity to specific visual marks.

A.5. Annotation File Details

All final annotations are stored in two JSON files, separated into temporal and spatial categories. The basic attributes include:

- `number`: the integer index identifier
- `views`: the image paths for each of the four synchronized views
- `question_type`: categorized as either spatial or temporal
- `question_subtype`: categorized as either coarse grained or fine grained
- `question_name`: the specific task name selected from the twelve benchmark categories
- `question`: the text prompt describing the question for the given images
- `answers`: representing the available options A through E, including one none of the above choice
- `gt_answer`: a single letter indicating the correct ground truth option

A.6. Benchmark License

CVBench is released strictly as a research benchmark intended for non commercial and academic use. The diverse human centric scenes are sourced from publicly available multiview datasets. We have rigorously reviewed and signed the respective data use agreements for all underlying source data, strictly adhering to original dataset requirements and ethical guidelines. We will not redistribute sensitive human biometric data, raw video files, or identifiable information without explicit and documented permission from the original dataset creators. The release of CVBench will be limited exclusively to the curated question and option pairs alongside the necessary metadata required to execute the evaluation protocols.

B. Detailed Comparison with Previous Benchmarks

The evaluation of modern visual and vision language models has been overwhelmingly predicated on single view scenarios. Consequently, high performance on existing benchmarks, which often rely on sufficient view assumptions, does not guarantee robustness in physical multiple camera environments. This evaluation paradigm rewards single view pattern recognition, leaving the capacity for cross view fusion largely unevaluated.

While many models accept multiple views as input, they are not methodically evaluated on their ability to synthesize complementary and often conflicting information. For instance, video centric benchmarks such as MMBench-Video [12], Video-MME [15], and LVBench [44] heavily focus on temporal reasoning within one continuous stream. However, this paradigm does not evaluate the ability to synthesize a coherent understanding from disparate viewpoints, leaving it indeterminate whether models perform true cross view fusion or simply cherry pick an optimal view.

A distinct line of inquiry challenges the single image constraint through benchmarks that test retrieval from large unordered image collections, such as Blink [16] and

Table 5. **Comprehensive Comparison with Existing Benchmarks.** **S**: Spatial Tasks; **T**: Temporal Tasks; **MV**: Multi-View; **Sync.**: Synchronization; **SVI**: Single-View Insufficiency. CVBench is uniquely positioned by simultaneously evaluating human centric and synchronized spatio temporal reasoning while strictly enforcing SVI across multiple granularities.

Benchmark	Core Focus	S	T	MV	Sync.	SVI	Human Centric	Granularity	Modality
Video & Temporal Context Benchmarks									
MMBench-Video [12]	Holistic Video Understanding	✗	✓	✗	✗	✗	✗	Coarse	Video
Video-MME [15]	Temporal Context Length	✗	✓	✗	✗	✗	✗	Coarse/Fine	Video
LVBench [44]	Long-Term Retrieval	✗	✓	✗	✗	✗	✗	Coarse	Video
Multi-Image & Visual Reasoning Benchmarks									
Blink [16]	Visual Perception	✓	✗	✗	✗	✗	✗	General	Images
MileBench [40]	Long-Context Retrieval	✓	✗	✗	✗	✗	✗	General	Images
MuirBench [42]	Multi-Image Relation	✓	✗	✓	✗	✗	✗	General	Images
AllAnglesBench [53]	Embodied Multi-view	✓	✗	✓	✓	✗	✗	General	Images
Medframeqa [56]	Clinical Reasoning	✓	✗	✗	✗	✗	✗	Fine	Medical
VLM2-Bench [59]	Visual Cues Linking	✓	✗	✗	✗	✗	✗	General	Images
CVBench (Ours)	**Cross View Human Underst.**	✓	✓	✓	✓	✓	✓	**Coarse & Fine**	**Multiple Videos & Images**

MileBench [40], or target domain specific relational reasoning, including Medframeqa [56]. While valuable, the inputs in these benchmarks are typically uncalibrated and non contemporaneous, precluding the evaluation of spatially consistent geometric or identity fusion.

CVBench is the **first** to integrate the synchronized data structure of classical multiview datasets with the language grounded evaluation paradigm of modern benchmarks. The primary distinction from all prior work is the systematic enforcement of verifiable single view insufficiency. Our benchmark design moves beyond permitting multiview input; it mandates multiview synthesis. A correct answer in CVBench cannot be reliably resolved by selecting an optimal single view. It must be synthesized from complementary, partial, and sometimes conflicting evidence. Table 5 summarizes these critical structural differences.

C. Experiments Details

C.1. Human Baseline Details

To establish a rigorous upper bound for human level spatial intelligence on our benchmark, we recruited eight independent postgraduate evaluators. These individuals were completely separated from the dataset curation process to prevent any prior exposure bias. The evaluators primarily possess academic backgrounds in computer science and biomechanics, equipping them with a strong foundational understanding of physical kinematics and spatial reasoning. To conduct the evaluation, participants were provided with a custom synchronized four view interface in Python Gradio equipped with comprehensive playback controls, allowing them to meticulously analyze temporal sequences and cross reference spatial alignments before making a decision. An average accuracy of $\sim 94\%$ was achieved, with low inter-evaluator variance, confirming that the tasks are unambiguously solvable when complementary multiple camera evidence is correctly synthesized by a capable reasoning agent.

C.2. Evaluation Setup Details

To ensure absolute reproducibility and deterministic outputs across all experiments, the generation temperature for all evaluated models is strictly set to zero, and a greedy decoding strategy is universally applied. During testing, we explicitly instruct the models to output only a single letter corresponding to their chosen answer. However, we observed that many models frequently generate conversational filler or extensive reasoning traces instead of adhering to the strict formatting constraint. To address this and ensure a perfectly fair comparison, we implement a two stage answer extraction pipeline. First, we apply a strict rule based filter designed to capture isolated letter predictions. If this initial filter fails to isolate a valid response, we employ a highly capable secondary model, specifically Gemini-2.5-Flash-Lite, to semantically parse the raw text and extract the final intended answer. Following established evaluation protocols from prior literature, we adopt this extraction method to guarantee fairness for smaller open weight models that might possess weaker instruction following capabilities.

C.3. Annotation Distribution

To strictly prevent models from exploiting statistical artifacts or developing shallow guessing heuristics, we shuffle the correct ground truth letter assignments to guarantee that the correct answers are perfectly and uniformly distributed among the five available options.

C.4. Details of Prompts

We present the complete structural details of our primary model evaluation prompt in Figure 3 alongside the specific answer extraction system prompt in Figure 4. Evaluated

models are explicitly instructed via the system prompt to analyze the synchronized visual evidence and output a single letter answer corresponding to their selected choice. For the secondary extraction phase, the input provided to the extraction model consists entirely of the raw text output generated by the benchmarked model. This extraction model is likewise instructed to output a single definitive letter. If the extraction model cannot definitively identify a clear selection from the provided raw text, it is required to output the specific word Invalid, at which point the evaluation script strictly marks the prediction as incorrect in our final accuracy calculations.

Prompt Template for CVBench Model Evaluation

System Prompt

```
You are an expert in multiple view human
understanding and three dimensional spatial
and temporal reasoning.  You must answer
strictly with one single letter from A, B, C,
D, or E. Do not add any explanation or extra
text.  Directly answer.
```

Input Template

```
You are given synchronized video clips of the
same scene.

Question:
{question}

Options:
A) {optA}
B) {optB}
C) {optC}
D) {optD}
E) {optE}

Output:  Answer with exactly ONE letter
from A, B, C, D, or E, for example, just "E".
No explanation.  Only one letter.
```

Figure 3. **Primary Evaluation Prompt.** This template instructs the benchmarked models to synthesize the synchronized visual evidence and output a single letter corresponding to their selected answer.

D. Additional Examples

We provide additional visualization examples in Figure 5, Figure 6, Figure 7, and Figure 8 to further demonstrate the benchmark tasks.

Prompt Template for CVBench Answer Extraction

System Prompt

```
You are an answer extractor.  You will be
given the raw text output from an artificial
intelligence model.  Your job is to identify
the final selected option from A, B, C, D,
or E. Output ONLY the single letter.  If
no clear letter is found, output the word
Invalid.
```

Figure 4. **Answer Extraction Prompt.** This template is utilized by the secondary extraction model to rigorously parse conversational or poorly formatted outputs from the primary models into a solitary letter.

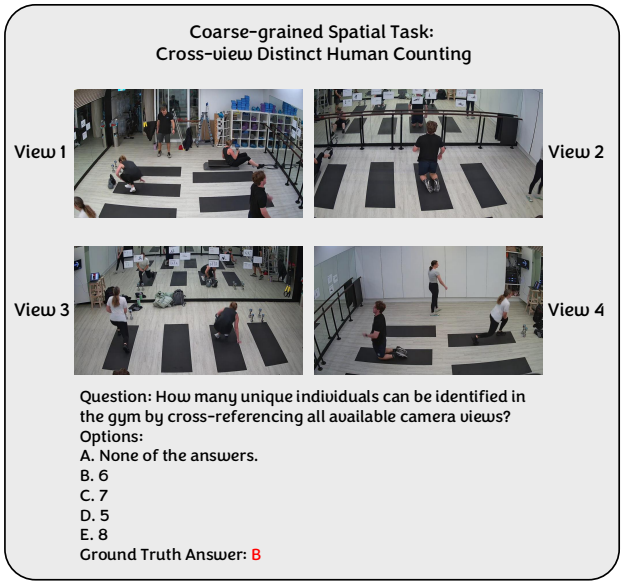


Figure 5. Additional Spatial Task Visualization Example 1.

Coarse-grained Spatial Task:
Cross-view Identify Association

Question: Which of the following identity matches across different images is correct?
Options:
A. Person 4 in the third image is the same individual as person 9 in the second image.
B. Person 6 in the third image is the same individual as person 13 in the first image.
C. Person 11 in the third image is the same individual as person 10 in the fourth image.
D. None of the answers.
E. Person 17 in the first image is the same individual as person 15 in the second image.
Ground Truth Answer: A

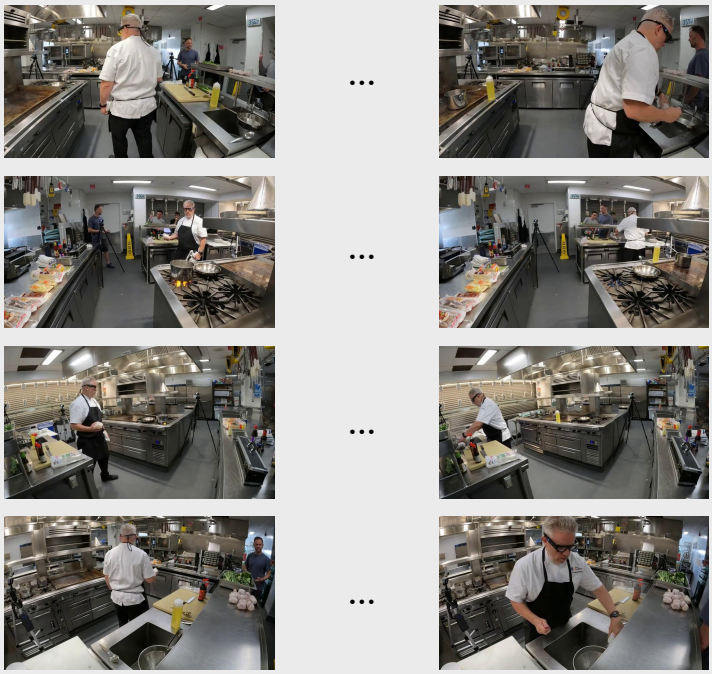
Figure 6. Additional Spatial Task Visualization Example 2.

Fine-grained Spatial Task:
Cross-view Body-Part Contact

Question: Which specific body part of the player nearest the ball is in physical contact with the soccer ball?
Options:
A. His left foot.
B. None of the answers.
C. His left knee.
D. His right foot.
E. His right ankle.
Ground Truth Answer: B

Figure 7. Additional Spatial Task Visualization Example 3.

Coarse-grained Temporal Task: Cross-view Motion Recognition



Question: Identify the primary activity being performed by the person in the foreground.
Options:
A. Cleaning and scrubbing a metal pot in a sink.
B. Slicing vegetables on a prep table near the stove.
C. Operating a digital control panel on an industrial oven.
D. Pour the contents of the metal pot into the sink.
E. None of the answers.
Ground Truth Answer: D

Figure 8. Additional Temporal Task Visualization Example 4.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[3] Chaitanya Bandi and Ulrike Thomas. Action recognition via multi-view perception feature tracking for human–robot interaction. *Robotics*, 14(4):53, 2025. 2

[4] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024. 2

[5] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5030–5039, 2018. 2, 3, 1

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[7] Jaeho Choi, Soheil Hor, Shubo Yang, and Amin Arbabian. Mvdoppler-pose: Multi-modal multi-view mmwave sensing for long-distance self-occluded human walking pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27750–27759, 2025. 1

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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. 6

[9] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S Lakshmipathy, Agniv Chatterjee, Michael J Black, and Dimitrios Tzionas. Pico: Reconstructing 3d people in contact with objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025. 4

[10] DeepSeek-AI. Deepseek-v3 technical report, 2024. 6

[11] Bohao Fan, Siqi Wang, Wenxuan Guo, Wenzhao Zheng, Jianjiang Feng, and Jie Zhou. Human-m3: A multi-view multi-modal dataset for 3d human pose estimation in outdoor scenes. *arXiv preprint arXiv:2308.00628*, 2023. 3, 1

[12] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 2, 3

[13] Wei Feng, Feifan Wang, Ruize Han, Yiyang Gan, Zekun Qian, Junhui Hou, and Song Wang. Unveiling the power of self-supervision for multi-view multi-human association and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[14] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024. 2

[15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2, 3

[16] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 3, 2

[17] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM international conference on multimedia*, pages 282–290, 2021. 3, 1

[18] Nikolaos Gkalelis, Hansung Kim, Adrian Hilton, Nikos Nikolaidis, and Ioannis Pitas. The i3dpost multi-view and 3d human action/interaction database. In *2009 Conference for Visual Media Production*, pages 159–168. IEEE, 2009. 1

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, 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*, pages 19383–19400, 2024. 3, 1

[20] Tianchen Guo, Heming Du, Huan Huo, Bo Liu, and Xin Yu. Who is being impersonated? deepfake audio detection and impersonated identification via extraction of id-specific features. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 301–320. Springer, 2024. 2

[21] Tianchen Guo, Peter Anthony Logan, Thomas Wackwitz, and David Martin. Plnet-12: A vision-language benchmark for zero-shot physical literacy analysis across 12 fundamental movements. In *Australasian Joint Conference on Artificial Intelligence*, pages 242–254. Springer, 2025. 2, 3

[22] Wenxuan Guo, Zhiyu Pan, Ziheng Xi, Alapati Tuerxun, Jianjiang Feng, and Jie Zhou. Sports analysis and vr viewing system based on player tracking and pose estimation with multimodal and multiview sensors. *arXiv preprint arXiv:2405.01112*, 2024. 2

[23] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language

models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8450–8460, 2025. 2

[24] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *ECCV*, 2020. 2, 3, 1

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[26] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, 83(5): 14885–14911, 2024. 2

[27] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19750–19762. IEEE, 2023. 3, 1

[28] Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Harry Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37:28798–28827, 2024. 2

[29] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 2

[30] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE Transactions on Multimedia*, 26:10015–10028, 2024.

[31] Chen Liu, Peike Patrick Li, Qingtao Yu, Hongwei Sheng, Dadong Wang, Lincheng Li, and Xin Yu. Benchmarking audio visual segmentation for long-untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22712–22722, 2024.

[32] Chen Liu, Peike Li, Liying Yang, Dadong Wang, Lincheng Li, and Xin Yu. Robust audio-visual segmentation via audio-guided visual convergent alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28922–28931, 2025.

[33] Chen Liu, Liying Yang, Peike Li, Dadong Wang, Lincheng Li, and Xin Yu. Dynamic derivation and elimination: Audio visual segmentation with enhanced audio semantics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3131–3141, 2025. 2

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 6

[35] Olivier Moliner. *Sparse Multi-View Computer Vision for 3D Human and Scene Understanding*. PhD thesis, Lund University, 2025. 1

[36] OpenAI. Introducing gpt-5, 2025. 6

[37] Shidong Pan, Tianchen Guo, Lihong Zhang, Pei Liu, Zhenchang Xing, and Xiaoyu Sun. A large-scale investigation of semantically incompatible apis behind compatibility issues in android apps. *arXiv preprint arXiv:2406.17431*, 2024. 2

[38] Feng Qiu, Heming Du, Wei Zhang, Chen Liu, Lincheng Li, Tianchen Guo, and Xin Yu. Learning transferable compound expressions from masked autoencoder pretraining. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4733–4741. IEEE, 2024.

[39] Feng Qiu, Wei Zhang, Chen Liu, Lincheng Li, Heming Du, Tianchen Guo, and Xin Yu. Language-guided multi-modal emotional mimicry intensity estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4742–4751. IEEE Computer Society, 2024. 2

[40] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024. 3

[41] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8001–8013, 2023. 4

[42] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 3

[43] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[44] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025. 2, 3

[45] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023. 2

[46] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 1

[47] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024. 1

[48] Qingzheng Xu, Huiqiang Chen, Heming Du, Hu Zhang, Szymon Łukasik, Tianqing Zhu, and Xin Yu. M3a: A multimodal misinformation dataset for media authenticity

analysis. *Computer Vision and Image Understanding*, 249: 104205, 2024. 2

[49] Qingzheng Xu, Ru Cao, Xin Shen, Heming Du, Sen Wang, and Xin Yu. M3gym: A large-scale multimodal multi-view multi-person pose dataset for fitness activity understanding in real-world settings. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12289–12300, 2025. 2, 3, 1

[50] Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. Mdam3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, pages 5285–5296, 2025. 2

[51] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 6

[52] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 5

[53] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. 1, 3

[54] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 2

[55] Yuanjiong Ying, Xian Huang, and Wei Dong. Multi-view active sensing for human–robot interaction via hierarchically connected tree. *Sensors and Actuators A: Physical*, 378: 115752, 2024. 2

[56] Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, et al. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025. 3

[57] Zhixuan Yu, Linguang Zhang, Yuanlu Xu, Chengcheng Tang, Luan Tran, Cem Keskin, and Hyun Soo Park. Multiview human body reconstruction from uncalibrated cameras. *Advances in Neural Information Processing Systems*, 35:7879–7891, 2022. 1

[58] Zaiyang Yu, Prayag Tiwari, Luyang Hou, Lusi Li, Weijun Li, Limin Jiang, and Xin Ning. Mv-reid: 3d multi-view transformation network for occluded person re-identification. *Knowledge-Based Systems*, 283:111200, 2024. 1

[59] Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7510–7545, 2025. 1, 3

[60] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge. *arXiv preprint arXiv:2403.10825*, 2024. 2

[61] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tianchen Guo, and Xin Yu. An effective ensemble learning framework for affective behaviour analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4761–4772. IEEE, 2024. 2

[62] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2

[63] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 3, 1

[64] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. *Advances in neural information processing systems*, 27, 2014. 1