

OMNITHOUGHTVIS: A Scalable Distillation Pipeline for Deployable Multimodal Reasoning Models

Yuanhao Yue, Chengyu Wang*, Yuanjie Lyu, Lei Shen, Jun Huang

Alibaba Group, Hangzhou, China

{yueyuanhao.yyh, chengyu.wcy, lyuyuanjie.lyj,
yuzhou.sl, huangjun.hj}@alibaba-inc.com

Abstract

Production multimodal systems must serve complex vision-language tasks under strict latency, memory, and cost constraints. Frontier multimodal large language models (MLLMs) deliver strong reasoning but are prohibitively expensive for online serving; compact alternatives fit deployment budgets but lack structured reasoning ability. We present OMNITHOUGHTVIS, a scalable distillation pipeline that converts open-source multimodal corpora into deployment-ready chain-of-thought (CoT) supervision for compact MLLMs. Starting from a diverse seed pool, the pipeline generates structured reasoning traces from a high-capacity teacher, jointly annotates reasoning difficulty, answer quality, and semantic task tags, and applies difficulty-aware filtering with tag-based diversity sampling to produce a reusable 1.8M-sample candidate pool. Practitioners construct budget-specific training subsets from this pool without re-running distillation. A key practical finding is that reasoning *difficulty* is a far more effective data selection signal than answer *quality*, which becomes uninformative when teacher outputs are already strong. Distilled Qwen3-VL models (2B–8B) consistently improve over undistilled baselines across nine benchmarks, with the distilled 4B model matching or exceeding the undistilled 8B on multiple tasks. In production, replacing a large-model serving pipeline with the distilled compact model reduces end-to-end response latency by 79% while maintaining comparable downstream quality.¹

1 Introduction

Production multimodal systems, powering visual question answering, content understanding, and user-facing decisions, must operate under tight latency, memory, and cost budgets. Multimodal large

language models (MLLMs) with chain-of-thought (CoT) reasoning (Bai et al., 2025b) achieve strong accuracy, but frontier-scale models are impractical for online serving. Compact models meet deployment constraints yet frequently fail on reasoning-intensive tasks (Hinton et al., 2015; Yang et al., 2025; Wang et al., 2025b). Closing this gap requires reasoning supervision *specifically* for compact model training under serving constraints.

A central bottleneck is the scarcity of multimodal data optimized for reasoning distillation. In the text domain, large-scale CoT corpora have accelerated open-model progress (Chung et al., 2024; Toshniwal et al., 2024; Yu et al., 2024; Cai et al., 2026). Open multimodal datasets, by contrast, remain dominated by instruction tuning or final-answer supervision (Liu et al., 2023; Chen et al., 2024a; Li et al., 2025), with limited support for structured reasoning transfer. Naïvely generating multimodal CoT data at scale introduces noise, redundancy, and task imbalance, making it difficult to construct training subsets that are both effective and efficient under fixed compute budgets.

Recent pipelines such as BEE (Zhang et al., 2026) and OpenMMReasoner (Zhang et al., 2025a) advance reproducible multimodal data curation. Our work complements these efforts with a deployment-oriented objective: converting large open corpora into a *reusable candidate pool* from which practitioners construct budget-specific training subsets for compact MLLMs. Rather than proposing novel individual techniques, we contribute a *validated end-to-end industrial recipe*—combining structured teacher generation, joint annotation, and diversity-aware subset selection—designed for repeatability and auditability in production data workflows.

Concretely, we present OMNITHOUGHTVIS. Starting from FineVision (Wiedmann et al., 2025) as a diverse seed pool, the pipeline generates structured CoT traces from a high-capacity teacher un-

* Corresponding author.

¹The resources are publicly available in the EasyDistill framework (Wang et al., 2025a).

der explicit format constraints, jointly annotates reasoning difficulty, answer quality, and semantic task tags, and applies three-stage quality control to produce a curated corpus of 1.8M samples. The resulting metadata enables controllable subset construction: practitioners select difficulty thresholds and tag coverage to match deployment constraints without re-running distillation.

We distill Qwen3-VL models from 2B to 8B and evaluate on nine multimodal benchmarks. The distilled models consistently outperform undistilled baselines, with the 4B model matching or exceeding the undistilled 8B on several tasks. A systematic comparison of selection signals reveals that reasoning *difficulty* is far more effective than answer *quality*—a practical finding that directly shaped our pipeline. In production, substituting the large-model pipeline with the distilled compact model reduces end-to-end response latency by 79% while maintaining comparable downstream quality. Our contributions are:

- **A validated industrial distillation recipe.** An end-to-end pipeline converting open multimodal corpora into deployment-ready reasoning supervision, designed for auditability and repeatability in production data workflows.
- **A budget-controllable corpus with actionable metadata.** A reusable 1.8M-sample candidate pool with difficulty, quality, and semantic tag annotations supporting flexible subset construction under varying compute and deployment budgets.
- **Production-grounded evaluation.** Consistent benchmark gains across model scales, together with a deployment case study demonstrating 79% latency reduction and maintained task quality in an online serving stack.

2 Related Work

Reasoning Distillation. Chain-of-thought prompting (Wei et al., 2022) shows that exposing intermediate reasoning steps improves complex decision-making. Follow-up work transfers this ability to smaller models via teacher-generated rationales (Hsieh et al., 2023; Ho et al., 2023; Mukherjee et al., 2023), while reinforcement learning with verifiable rewards further improves frontier-model reasoning (DeepSeek-AI, 2025). Large text-domain reasoning corpora (Guha et al., 2026; Cai et al., 2026) accelerate open-model progress. We

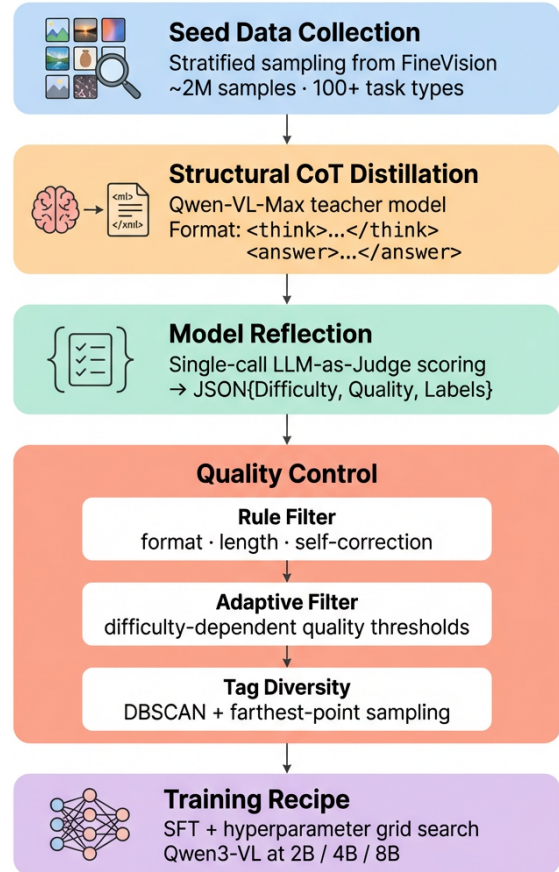


Figure 1: Overview of our pipeline.

extend these ideas to multimodal reasoning with a focus on *deployment-oriented* supervision for compact, latency-constrained MLLMs.

Multimodal Data Curation. Visual instruction tuning (Liu et al., 2023) and scaled captioning (Chen et al., 2024a) improve vision-language capabilities but target short-form or perceptual supervision rather than multi-step reasoning. Aggregation efforts such as Cambrian-1 (Tong et al., 2024) broaden coverage without specifically targeting reasoning distillation. Recent pipelines including BEE (Zhang et al., 2026) and OpenMMReasoner (Zhang et al., 2025a) advance reproducible multimodal data curation at scale. Our work differs in *objective* rather than technique: instead of building a fixed training set, we treat open corpora as a *seed pool* and produce a metadata-rich candidate corpus annotated with difficulty, quality, and semantic tags. This reusable-pool design and the accompanying difficulty-aware selection strategy address an operational need that prior pipelines do not explicitly target.

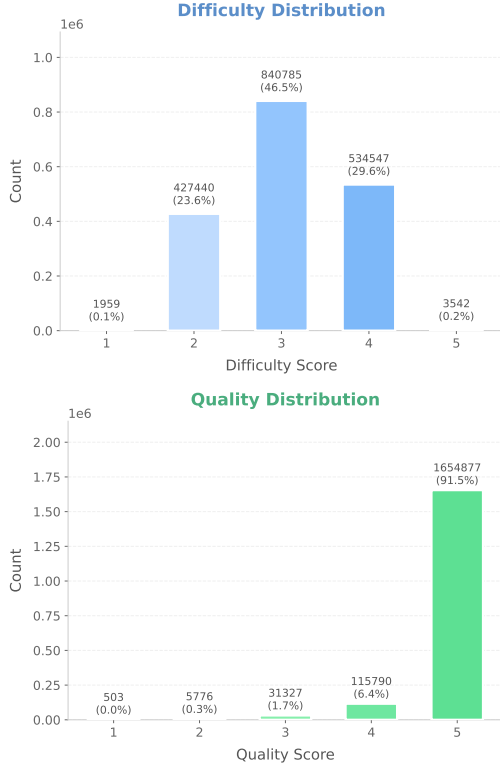


Figure 3: Difficulty and quality distributions. Difficulty spans a broad range; quality concentrates at high scores, motivating difficulty-first selection.

prove reasoning transfer than quality-based filtering, which is uninformative when teacher output quality is uniformly high.

Stage 3 (Tag-Based Diversity Sampling): To reduce over-representation of frequent categories, we perform diversity-aware selection via semantic tags. Tags are embedded with Qwen3-Embedding (Zhang et al., 2025b), near-synonyms (e.g., spatial/position) are merged with DBSCAN (Ester et al., 1996) ($\epsilon=0.3$, $\text{min_samples}=2$), and each sample is represented by its mean clustered-tag embedding. Farthest-point sampling in this space produces a more diverse subset. The tag space also enables domain-targeted curation: practitioners filter by tags for specific deployment needs without re-running the full pipeline.

Candidate pool design: The full corpus contains 1.8M curated samples, designed as a *reusable resource*: practitioners select budget-specific subsets by adjusting difficulty thresholds and tag coverage. This curate-once, select-many-times workflow amortizes the cost of teacher generation across training configurations.

3.5 Dataset Statistics

Figure 3 summarizes key distributions. Difficulty is approximately unimodal, centered at Level 3 (46.5%), with Levels 2 (23.6%) and 4 (29.6%) well represented; extreme levels are rare. Quality is strongly right-skewed (Level 5: 91.5%), confirming its limited discriminative value. Semantic tags show broad coverage: reasoning (85%), object (75%), scene (45%), count (38%), text (37%), math (33%).

4 Experiments

4.1 Setup

Training. We use Qwen3-VL-2B, -4B, and -8B as student backbones (distilled models: DistilQwen3-VL). All models are trained on a 0.5M-sample subset selected from the 1.8M candidate pool via difficulty-aware selection (§3.4). Training uses AdamW ($\beta_1=0.9$, $\beta_2=0.999$, weight decay $=0.01$), batch size 128, sequence length 3072, 3% warmup, and cosine decay to 10% of peak learning rate. Learning rate and epoch count are tuned on a held-out split (§4.3).

Benchmarks. Nine benchmarks spanning general understanding and mathematical reasoning: AI2D (Kembhavi et al., 2016), MMBench (Liu et al., 2024), MMStar (Chen et al., 2024b), MMMU (Yue et al., 2024), MMMU-Pro (Standard & Vision) (Yue et al., 2025), MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), and MathVision (Wang et al., 2024).

Evaluation. Models use thinking prompts (Appendix A.1) with temperature 0.5. Multiple-choice benchmarks report exact-match accuracy from `<answer>` tags. Open-ended tasks use Qwen3-VL-Plus as the judge model; these results serve as secondary evidence.

Model family alignment. Our pipeline uses Qwen-family models throughout: teacher (Qwen-VL-Max), scorer (Qwen3-VL-Flash), students (Qwen3-VL), and judge (Qwen3-VL-Plus). This reflects a pragmatic production choice: these models are the primary infrastructure in our deployment environment. Importantly, 7 of 9 benchmarks use exact-match accuracy, which is invariant to model-family stylistic alignment.

4.2 Main Results

Table 1 shows that distillation consistently improves performance across all scales. Gains are

Model	General		Knowledge & Understanding				Mathematical Reasoning		
	MMB	MMStar	AI2D	MMMU _{val}	MMMU _{Pro-S}	MMMU _{Pro-V}	MathVista	MathVerse	MathVision
Qwen3-VL-2B	74.3	56.6	71.3	47.9	32.1	27.5	60.0	34.1	20.4
DistilQwen3-VL-2B	74.7	61.5	75.4	49.2	35.6	31.9	65.0	43.0	27.3
Qwen3-VL-4B	82.0	66.6	79.4	56.7	41.3	40.3	70.7	36.4	31.9
DistilQwen3-VL-4B	82.7	68.4	81.6	60.3	46.9	45.2	74.0	53.2	32.9
Qwen3-VL-8B	82.6	69.3	81.0	60.0	46.9	45.3	72.5	38.2	35.2
DistilQwen3-VL-8B	84.7	71.1	83.7	62.8	50.1	48.6	74.8	55.5	40.1

Table 1: Benchmark results. MMB = MMBench, Pro-S/V = MMMU-Pro Standard/Vision. All values are accuracy (%). Multiple-choice benchmarks (7 of 9) use exact-match; open-ended tasks use model-based judging as a secondary signal.

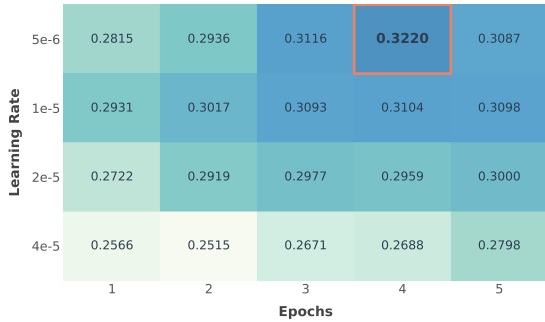


Figure 4: Hyperparameter grid on MMMU_{Pro-V} for DistilQwen3-VL-2B.

largest on reasoning-intensive benchmarks: MathVerse (+8.9 to +17.3 points) and MMMU-Pro (+3.3 to +5.6). Notably, the distilled 4B model matches or exceeds the undistilled 8B on AI2D (81.6 vs. 81.0) and MMMU (60.3 vs. 60.0), indicating that structured reasoning supervision can partially compensate for model scale, which is a favorable trade-off for deployment.

4.3 Analysis

Hyperparameter tuning. We evaluate four learning rates (5×10^{-6} to 4×10^{-5}) across 1–5 epochs for DistilQwen3-VL-2B using MMMU-Pro Vision (Figure 4). The best setting is: 5×10^{-6} for 4 epochs. Higher learning rates consistently degrade performance, and training beyond 4 epochs provides no stable gains.

Data selection signals. Under a fixed budget ($\sim 100K$ samples), we compare filtering strategies (Table 2). Quality-only filtering (≥ 5) *hurts* relative to random sampling (0.2948 vs. 0.3093), confirming that quality scores are uninformative when teacher output quality is uniformly high. Difficulty-only filtering (≥ 4) performs best (0.3156). This result directly informed our pipeline: difficulty is the primary selection signal, quality serves only as

Filtering Strategy	MMMU _{Pro-V}
No filtering (random 100K)	0.3093
Quality ≥ 5 only	0.2948
Difficulty ≥ 4 only	0.3156
Quality ≥ 5 & Difficulty ≥ 4	0.3139

Table 2: Filtering strategy ablation ($\sim 100K$ samples each). Results show that difficulty is a more effective selection signal than quality.

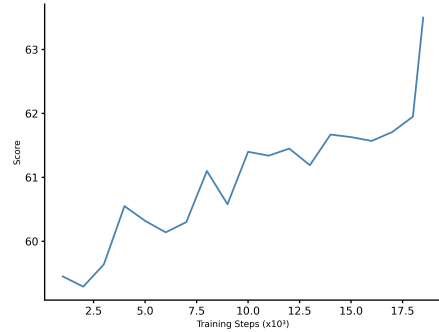


Figure 5: Average nine-benchmark performance during DistilQwen3-VL-8B training.

a coarse defect filter.

Training dynamics. Figure 5 tracks DistilQwen3-VL-8B over training. The average score rises from 59.5 (1K steps) to 63.5 (19K steps). Per-benchmark curves (Figure 6) reveal that general visual benchmarks (AI2D, MMBench) plateau early, while reasoning tasks (MathVerse, MMMU-Pro) continue improving, suggesting that staged training allocation may better utilize fixed compute budgets.

5 Deployment and Discussion

5.1 Production Case Study

We evaluate the distilled model in a production multimedia intelligence serving stack (Figure 7). The architecture places a compact MLLM on the online critical path: it processes user queries, interacts with an internal multimodal knowledge base

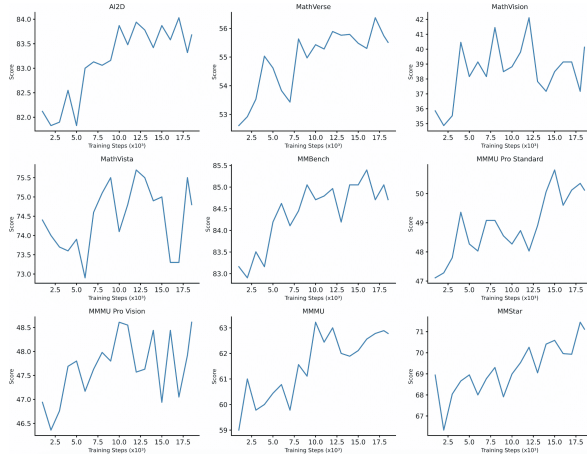


Figure 6: Per-benchmark scaling. General benchmarks saturate; reasoning tasks benefit from longer training.

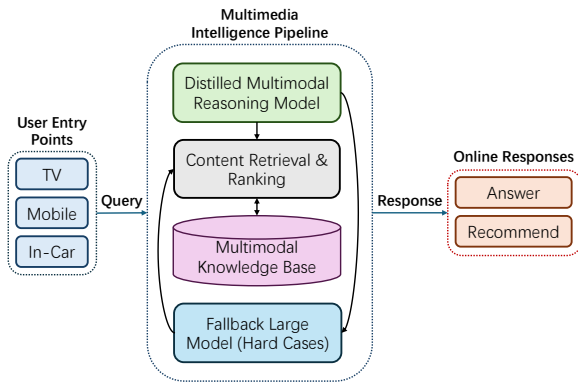


Figure 7: Production multimedia intelligence stack. A compact MLLM handles online requests; a larger model serves as hard-case fallback.

for retrieval and ranking, and generates user-facing responses. A larger fallback model handles cases flagged as hard.

We run an internal A/B test comparing the baseline pipeline (large model + online search) against our redesigned pipeline (distilled compact model + knowledge base retrieval), under the same serving chain and workload:

- (1) **Maintained downstream quality.** On the deployed user-intent classification task, the baseline achieves 92% task quality; the distilled compact model maintains comparable quality (slightly above 90%). Reasoning-focused distillation preserves the decision quality required for production.
- (2) **Improved retrieval accuracy.** The redesigned pipeline, which pairs the distilled model’s structured query understanding with retrieval from the internal knowledge base, improves retrieval accuracy by 58% relative to the baseline pipeline (large model with online search).

- (3) **Reduced latency.** The distilled pipeline reduces average end-to-end response latency by 79% on production traffic, which is the most impactful operational gain.

Together, these results demonstrate that the distilled compact model achieves a substantially better quality–efficiency trade-off than the baseline large-model pipeline, validating the practical value of large-scale reasoning distillation for applications.

5.2 Practical Insights

We distill several lessons from building and deploying this pipeline:

Invest in difficulty estimation, not quality rubrics. When the teacher is sufficiently capable, quality scores collapse to the high end and carry little discriminative signal. Difficulty provides a more actionable distribution. Teams building similar pipelines should prioritize difficulty estimation.

Budget training time by capability type. General visual benchmarks plateau early; reasoning tasks continue improving with longer training. Under compute budgets, staged curricula that allocate more optimization steps to reasoning-intensive data may outperform uniform schedules.

Build reusable candidate pools. Although we train on 0.5M samples, the full 1.8M pool supports diverse downstream use cases (e.g., math-focused vs. general-purpose models) via tag-based re-sampling, amortizing the one-time cost of teacher generation across deployment scenarios.

Within-family distillation: a pragmatic trade-off. Using Qwen-family models throughout is a deliberate choice reflecting our production ecosystem, consistent with common industrial practice. It introduces the risk that some gains partly reflect stylistic alignment. We mitigate this by relying on exact-match evaluation for the majority of benchmarks (7 of 9) and treating model-judged results as secondary evidence. Cross-family validation remains a valuable direction for future work.

6 Conclusion

We presented OMNITHOUGHTVIS, a scalable distillation pipeline that converts open multimodal corpora into deployment-ready reasoning supervision. The pipeline produces a reusable 1.8M-sample candidate pool with difficulty, quality, and semantic tag metadata, enabling budget-specific subset construction without re-running distillation. Distilled Qwen3-VL models consistently outper-

form undistilled baselines across nine benchmarks, and a production deployment achieves 79% latency reduction with comparable task quality. Our central practical finding, reasoning difficulty is far more effective than answer quality as a data selection signal, offers actionable guidance for practitioners building similar pipelines.

Limitations

Our pipeline uses Qwen-family models at all stages (teacher, scorer, student, judge), which is a pragmatic choice reflecting our production infrastructure. To ensure this does not inflate results, we rely on exact-match accuracy for 7 of 9 benchmarks, a metric that is invariant to model-family stylistic alignment. Extending to cross-family students (e.g., LLaMA, InternVL) is a natural next step that we plan to explore.

Quality scores in our corpus are concentrated at high levels (91.5% at Level 5), which limits their discriminative value for subset selection. Rather than treating this as a pipeline defect, we surface it as a practical finding: when the teacher is sufficiently capable, difficulty is the more actionable curation signal. Our ablation validates this conclusion quantitatively (§4.3).

Ethical Considerations

OMNITHOUGHTVIS is constructed from publicly available, open-source vision-language datasets using automated generation, scoring, and filtering. The dataset may inherit biases from source data, teacher models, or scoring models. Because annotations are largely automated, some samples may contain inaccurate reasoning traces. The dataset is primarily English-language and vision-centric, which may limit coverage across languages and populations. All source datasets were checked for license compatibility, and we only include data intended for research use. We encourage responsible use and additional evaluation before applying models in downstream settings.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang

Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025a. [Qwen3-vl technical report](#).

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025b. [Qwen2.5-vl technical report](#).

Wenrui Cai, Chengyu Wang, Junbing Yan, Jun Huang, and Xiangzhong Fang. 2026. [Reasoning with omnitought: A large cot dataset with verbosity and cognitive difficulty annotations](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pages 8431–8450. Association for Computational Linguistics.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. [Sharegpt4v: Improving large multi-modal models with better captions](#). In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVII*, pages 370–387, Berlin, Heidelberg. Springer-Verlag.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. [Are we on the right way for evaluating large vision-language models?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

DeepSeek-AI. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, pages 226–231. AAAI Press.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer,

- Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanxia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. 2026. [Openthoughts: Data recipes for reasoning models](#). In *The Fourteenth International Conference on Learning Representations*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. [Large language models are reasoning teachers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [LLaVA-onevision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, pages 216–233, Berlin, Heidelberg. Springer-Verlag.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations*.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.
- Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, Adithya Jairam Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. [Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *Advances in Neural Information Processing Systems*, 37:34737–34774.
- Chengyu Wang, Junbing Yan, Wenrui Cai, Yuanhao Yue, and Jun Huang. 2025a. [Easydistill: A comprehensive toolkit for effective knowledge distillation of large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 787–795. Association for Computational Linguistics.
- Chengyu Wang, Junbing Yan, Yuanhao Yue, and Jun Huang. 2025b. [Distilqwen2.5: Industrial practices of training distilled open lightweight language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 32–42. Association for Computational Linguistics.
- Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. 2025. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*.
- Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2025. [Survey on knowledge distillation for large language models: Methods, evaluation, and application](#). *ACM Trans. Intell. Syst. Technol.*, 16(6).

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Metamath: Bootstrap your own mathematical questions for large language models](#). In *The Twelfth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2025. [MMM-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.

Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Ziwei Liu, Xingxuan Li, and Lidong Bing. 2025a. Openmmreasoner: Pushing the frontiers for multimodal reasoning with an open and general recipe. *arXiv preprint arXiv:2511.16334*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. [Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?](#) In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, pages 169–186, Berlin, Heidelberg. Springer-Verlag.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Hengrui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi min Hu. 2026. [Bee: A high-quality corpus and full-stack suite to unlock advanced fully open MLLMs](#). In *The Fourteenth International Conference on Learning Representations*.

A Technical Details

A.1 Distillation Prompt Template

System Prompt.

You are a helpful assistant to think step by step. Provide your reasoning steps within

<think></think> tags and give your final answer within <answer></answer> tags.

Query Structure. Given an image and a question, the teacher input follows:

```
### Image
<image>
### Question
{question}
### Output Format
(Strictly Enforced)
<think>
Clearly explain your reasoning step by step. Describe how you arrived at the conclusion. The reasoning process MUST BE enclosed within <think> </think> tags.
/<think>
<answer>
Your final answer to the user's question.
</answer>
```

A.2 Automatic Scoring Protocol

We use Qwen3-VL-Flash as scorer, evaluating each sample on reasoning difficulty and annotation quality.

Reasoning Difficulty. $d_i \in \{1, 2, 3, 4, 5\}$:

- **1 (Very Easy):** Simple color or shape identification.
- **2 (Easy):** Basic counting or straightforward spatial relationships.
- **3 (Moderate):** Brief reasoning or common action/attribute identification.
- **4 (Hard):** Multi-step reasoning, subtle visual cues, or rare concepts.
- **5 (Very Hard):** Abstract reasoning, complex scene understanding, or high ambiguity.

Annotation Quality. $q_i \in \{1, 2, 3, 4, 5\}$:

- **1 (Very Low):** Entirely incorrect or irrelevant.
- **2 (Low):** Mostly incorrect; minor correct elements.

- **3 (Medium):** Partially correct; key details missing.
- **4 (High):** Largely correct; minor inaccuracies.
- **5 (Very High):** Fully accurate and complete.

As noted in §3.3, quality is strongly right-skewed (91.5% at Level 5) and less useful than difficulty for subset selection.

A.3 Semantic Tag Definitions

Each sample receives one or more semantic tags characterizing its reasoning domain, visual theme, and task type. Coverage includes:

- **Visual Understanding:** object, attribute, scene.
- **Spatial Reasoning:** spatial, position, layout, direction, coordinate.
- **Chart/Table Interpretation:** chart, graph, table, bar chart, pie chart, scatter plot, axis, legend.
- **Logical/Mathematical Reasoning:** count, comparison, math, algebra, geometry, percentage, ratio.
- **Domain Knowledge:** physics, chemistry, biology, code, algorithm, history.
- **Reflective Reasoning:** verification, fact-check, error, discrepancy.

B Dataset-Level Case Studies

We compare FineVision source annotations with OMNITHOUGHTVIS structured annotations, showing how the pipeline transforms short-form supervision into explicit, auditable reasoning traces.

Case 1 (Scientific concept grounding). FineVision provides a brief justification for “ionic” bonding. Our annotation follows a structured checklist—entity identification → bonding analysis → dominance judgment → structural verification—tying the answer to observable cues.

Case 2 (Geometry; error correction). FineVision outputs option A without verification. Our annotation decomposes into geometry subgoals and recovers the correct answer (D).

Case 3 (Meme affect analysis). FineVision gives coarse labels. Our annotation allocates one

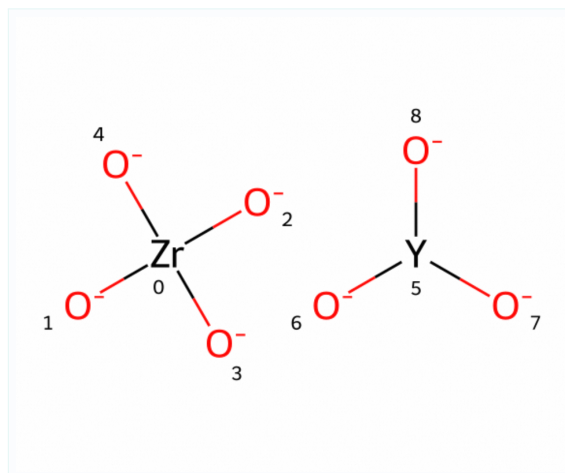
reasoning step per dimension (humor, sarcasm, offensiveness, motivation, sentiment), grounded in textual and visual evidence.

Case 4 (Chart reasoning; error correction). FineVision answers “Yes” without comparison. Our annotation performs curve-by-curve analysis, arriving at the correct answer (No).

Case 5 (Visual search in a cluttered scene). FineVision provides a brief description. Our annotation uses enumerate-then-identify: listing all individuals and their attributes before identifying the exception.

Full examples: Figures 8–12.

Case 1



Question: Does this compound exhibit ionic or covalent bonding?

■ **Original Annotation (FineVision):** The interactions between zirconium and oxygen in this structure indicate ionic bonding characteristics, particularly due to the presence of charged oxygen. **Answer:** ionic

★ Structured Annotation (Ours):

<think>

To determine whether the compound exhibits ionic or covalent bonding, we need to analyze the nature of the bonds between the atoms in the structure.

Step 1: Identify the elements involved. The structure contains zirconium (Zr), yttrium (Y), and oxygen (O). Oxygen is highly electronegative, while Zr and Y are metals.

Step 2: Analyze metal–oxygen bonding. Zr and Y are transition metals bonded to oxygen atoms. Oxygen typically forms anions (O^-) in such compounds. The presence of O^- ions suggests electrons are transferred from the metals to oxygen, forming ionic bonds.

Step 3: Evaluate the dominant bonding type. In oxides of metals like Zr and Y, the bonding is primarily ionic due to the large difference in electronegativity. Although some covalent character may exist due to polarization, the dominant bonding type remains ionic.

Step 4: Structural verification. The structure shows Zr and Y each bonded to four O^- ions, typical for ionic oxides. The arrangement resembles a perovskite-like structure with metal cations surrounded by oxygen anions, consistent with ionic bonding.

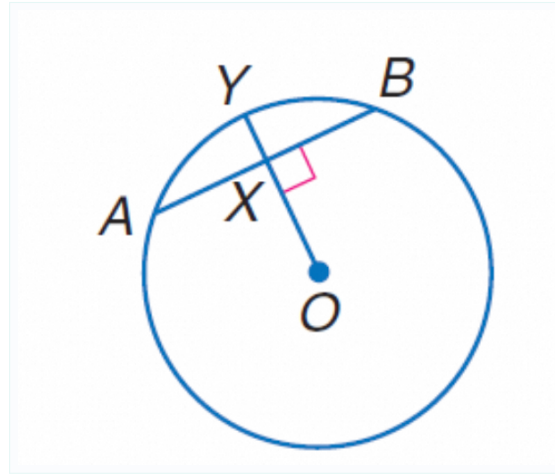
Conclusion: Given the involvement of metals (Zr, Y) and oxygen, and the presence of O^- ions, the bonding is predominantly ionic.

</think>

<answer> Ionic bonding </answer>

Figure 8: Case 1: Scientific concept grounding.

Case 2



Question: Answer the mathematical geometry problem and directly provide the correct option letter.

Circle O has a radius of 10, $AB = 10$ and $m\widehat{AB} = 60$. Find $m\widehat{AX}$.

Choices: A: 5 B: 10 C: 20 D: 30

■ **Original Annotation (FineVision):** A

★ Structured Annotation (Ours):

<think>

We are given: Circle O has a radius of 10, $AB = 10$, and $m\widehat{AB} = 60^\circ$. We need to find $m\widehat{AX}$.

Step 1: Analyze triangle OAB . Since AB is a chord and $OA = OB = 10$ (radii), and $AB = 10$, triangle OAB has all sides equal to 10 — it is an equilateral triangle. Therefore $\angle AOB = 60^\circ$, consistent with $m\widehat{AB} = 60^\circ$.

Step 2: Identify point X . Point X is the foot of the perpendicular from center O to chord AB . By the perpendicular-from-center theorem, $OX \perp AB$ bisects the chord, so $AX = XB = AB/2 = 5$.

Step 3: Compute the arc. Since OX bisects both chord AB and central angle $\angle AOB$, we have $\angle AOX = \angle BOX = 60^\circ/2 = 30^\circ$. Arc AX corresponds to central angle $\angle AOX$, so $m\widehat{AX} = 30^\circ$.

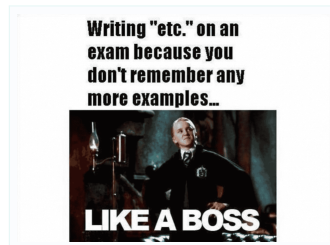
Conclusion: $m\widehat{AX} = 30$. The correct answer is **D**.

</think>

<answer>D </answer>

Figure 9: Case 2: Geometry with error correction (FineVision: A; correct: D).

Case 3



Question: Break down this meme’s emotional profile: humor level, sarcasm, offensiveness, motivational qualities, and sentiment.

■ **Original Annotation (FineVision):** This meme is funny. It is generally sarcastic, not offensive, and not motivational. The overall sentiment is positive.

★ **Structured Annotation (Ours):**

<think>

To analyze the emotional profile of this meme, let’s break it down step by step:

1. Humor Level. The humor stems from relatability and irony. Many students have experienced the frustration of being unable to recall examples during an exam and resorting to writing “etc.” as a placeholder. The image of Draco Malfoy, known for his smug demeanor, exaggerates this behavior. This contrast between a common academic struggle and a character known for superiority creates a light-hearted, situational humor.

2. Sarcasm. There is a strong element of sarcasm. The phrase “LIKE A BOSS” is typically used to mock someone acting overly confident. Here, it is applied to a student who pretends to be knowledgeable by using “etc.,” highlighting the absurdity of feigning expertise.

3. Offensiveness. The meme is not offensive. It does not target any specific group, individual, or sensitive topic. Instead, it pokes fun at a universal student experience in a non-malicious way.

4. Motivational Qualities. The meme lacks motivational qualities. It mocks a common academic shortcut rather than inspiring positive behavior, making the audience laugh at their own past struggles.

5. Sentiment. The overall sentiment is negative but in a playful, ironic way. It expresses frustration with academic limitations through humor and exaggeration. The tone is self-deprecating and slightly mocking, but not mean-spirited.

Summary: The meme uses relatable academic anxiety, exaggerated confidence, and sarcasm to create humor without crossing into offensiveness.

</think>

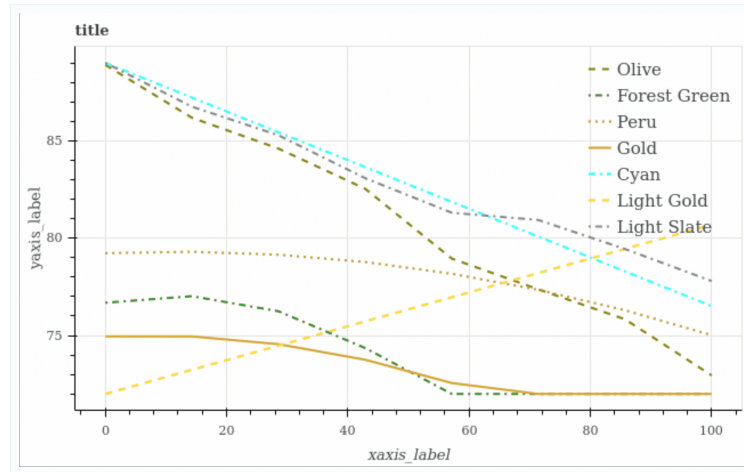
<answer>

Humor Level: High (relatable, ironic, and exaggerated); Sarcasm: High (mocking the pretense of knowledge); Offensiveness: Low (no targeted insults or sensitive topics); Motivational Qualities: None (mocks rather than inspires); Sentiment: Playful negativity (self-deprecating humor with irony)

</answer>

Figure 10: Case 3: Meme affect analysis with multi-dimensional reasoning.

Case 4



Question: Does Forest Green have the maximum area under the curve? Answer yes or no.

■ **Original Annotation (FineVision):** Yes.

★ **Structured Annotation (Ours):**

<think>

To determine whether Forest Green has the maximum area under the curve, we need to compare the areas under all lines in the graph. The area under a curve is proportional to how high the line stays above the x -axis over the entire range.

Step 1: Analyze each line. **Cyan** starts at the highest value (~ 98) and decreases steadily but remains relatively high throughout. **Light Slate** also starts high and decreases gradually, staying above most other lines except Cyan. **Olive** starts high but drops sharply after the first few points. **Peru** remains fairly flat around 78–79. **Gold** starts low and increases slowly. **Light Gold** starts low and increases steadily, crossing some lines but not reaching the top. **Forest Green** starts around 75, peaks slightly, then drops sharply and ends near zero.

Step 2: Compare areas. Cyan and Light Slate are the two highest lines throughout the range. Cyan starts higher and maintains a higher position for most of the range. Forest Green, while starting moderately high, drops quickly and ends near zero, so its area is significantly smaller.

Conclusion: Forest Green does **not** have the maximum area under the curve. Cyan has the largest area due to its consistently high values across the entire range.

</think>

<answer>No </answer>

Figure 11: Case 4: Chart reasoning with error correction (FineVision: Yes; correct: No).

Case 5



Question: Point out the person in the picture who is not wearing a helmet.

■ **Original Annotation (FineVision):** In the image, there is a person wearing a pink shirt and dark pants, standing near the center of the image to the right of a blue truck, who is not wearing a helmet. This individual is in contrast to the others around who are wearing helmets, presumably for safety reasons at this construction or work site.

★ Structured Annotation (Ours):

<think>

To determine who is not wearing a helmet, I will examine each person in the image:

Step 1: Enumerate all individuals.

- (1) Person on the left side of the blue truck: orange outfit, blue helmet.
- (2) Person next to them: orange outfit, blue helmet.
- (3) Person in the center-left: red outfit, blue helmet.
- (4) Person in the center-right: dark outfit, blue helmet.
- (5) Person near the bottom center: gray shirt, black pants, **no helmet**; holding a measuring tape.
- (6) Person near the right side under the makeshift shelter: blue helmet.
- (7) Person near the far right, partially obscured: blue helmet.

Step 2: Identify the exception.

After reviewing all individuals, only person (5) in the gray shirt is not wearing a helmet.

</think>

<answer>

The person in the gray shirt near the bottom center of the image is not wearing a helmet.

</answer>

Figure 12: Case 5: Fine-grained visual search in a cluttered scene.