Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Chinese Dataset for Training AI Models to Assess Procedural Accuracy and Scientific Compliance in VirtualMedical Simulation

Chaoya Wang

wcy_2003@163.com

Anhui Health College

Li Zhang

West Anhui Health Vocational College

data-descriptor

Keywords: virtual simulation, medical education, artificial intelligence, procedural accuracy, skill assessment, dataset

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8691386/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Virtual medical simulation has emerged as an indispensable pedagogical modality in health professions education, particularly in resource-constrained settings where clinical access remains limited and standardized patient resources are scarce. However, the challenge of ensuring procedural accuracy and scientific compliance during simulation-based learning through automated assessment remains inadequately addressed. This paper presents the Virtual Medical Simulation Training (VMST) dataset, a comprehensive Chinese-language corpus specifically designed to train artificial intelligence models for identifying and evaluating procedural accuracy and standardization in virtual medical simulation environments. The VMST dataset comprises 15,847 annotated operation sequences collected from 2,156 vocational medical students across 12 simulation training modules, encompassing nursing procedures, clinical examination techniques, and emergency response protocols. Each operation sequence includes timestamped action logs, expert-annotated accuracy labels, compliance scores based on standardized protocols, and detailed error classifications. Technical validation demonstrates that deep learning models trained on VMST achieve 89.3% accuracy in identifying procedural errors and 91.7% precision in compliance assessment, substantially outperforming baseline methods. This dataset addresses a critical gap in AI-assisted medical education assessment and establishes a foundation for developing intelligent tutoring systems capable of providing real-time feedback to students in virtual simulation environments.

Background & Summary

Healthcare systems worldwide confront unprecedented challenges in producing sufficient qualified health professionals to meet escalating demands for medical services. According to the World Health Organization, developing countries require an additional 18 million health workers by 2030 to achieve universal health coverage[1]. This workforce shortage is particularly acute in regions where traditional clinical training resources remain scarce, characterized by limited access to teaching hospitals, insufficient numbers of qualified clinical preceptors, and inadequate opportunities for direct patient contact[2]. The advent of virtual medical simulation technology offers a transformative solution to these educational constraints by providing repeatable, safe, and standardized training environments that transcend geographical and resource limitations[3].

The pedagogical advantages of virtual simulation in medical education extend far beyond accessibility. Simulation-based training enables students to practice complex procedures without risk to actual patients, receive immediate performance feedback, engage in unlimited repetition until competency is achieved, and experience standardized assessment conditions that minimize evaluator bias[4,5]. Evidence indicates that simulation training, when properly implemented, can reduce medical errors by up to 50%[6], with students who receive simulation-based education demonstrating superior clinical performance compared to those trained through traditional methods alone[7]. The COVID-19 pandemic further

accelerated the adoption of virtual simulation technologies as educational institutions globally sought alternatives to in-person clinical rotations[8].

Vocational medical education in China faces distinctive challenges. On one hand, institutions contend with high student-to-faculty ratios, substantial equipment costs, and significant clinical risks; on the other, they struggle with difficulty accessing clinical training sites, challenges in standardizing competency assessment, and obstacles to providing individualized feedback[9]. Chinese vocational medical colleges annually enroll over 500,000 students in nursing, medical imaging, medical laboratory technology, and allied health programs, yet many institutions maintain clinical training station-to-student ratios exceeding 1:20[10]. This structural imbalance results in students receiving an average of fewer than 40 hours of supervised hands-on practice throughout their entire educational program, substantially below the internationally recommended minimum of 200 hours[11].

In response to these resource constraints, virtual simulation platforms have been increasingly deployed across Chinese medical education institutions. These platforms generate substantial quantities of student interaction data, including operation sequences, temporal information, and outcome metrics. However, the potential value of these data for automated assessment and intelligent tutoring remains largely untapped, primarily due to the absence of high-quality annotated datasets suitable for training machine learning models[12]. As illustrated in Table 1, existing educational datasets in the medical domain predominantly focus on question-answering, diagnostic reasoning, or clinical note generation, with relatively limited attention to procedural skill assessment in simulation environments.

Table 1
Comparison of existing medical education datasets with VMST.

Dataset	Size	Language	Data Type	Task	Procedural Annotation
MedQA[13]	61,097	EN/ZH	Text QA	Knowledge assessment	No
PubMedQA[14]	1,000	EN	Text QA	Research reasoning	No
MedMCQA[15]	194,000	EN	MCQ	Medical examination	No
OULAD[16]	32,593	EN	Learning logs	Learning analytics	No
SciQ[17]	13,679	EN	Text QA	Science education	No
VMST (this study)	15,847	ZH	Operation sequences	Accuracy detection	Yes

EN = English, ZH = Chinese, MCQ = Multiple Choice Questions, QA = Question Answering.

The integration of artificial intelligence with medical education assessment represents a paradigm shift from traditional evaluation methodologies toward automated evaluation. Conventional procedural skill assessment relies on direct observation by trained evaluators using standardized checklists, a process that is time-intensive, subject to inter-rater variability, and challenging to implement at scale[18]. AI-based assessment systems hold the potential to provide continuous, objective, and scalable evaluation that can deliver immediate feedback to learners, identify specific areas requiring remediation, and track skill development over time[19]. However, development of such systems requires training data that captures the full spectrum of student performance, from novice errors to expert-level execution, accompanied by detailed annotations that enable machine learning algorithms to learn discriminative features between correct and incorrect procedures.

To address this critical gap, we present the Virtual Medical Simulation Training (VMST) dataset, the first large-scale Chinese-language dataset specifically designed for training AI models to assess procedural accuracy and scientific compliance in virtual medical simulation environments. The dataset comprises 15,847 complete operation sequences collected from vocational medical students performing standardized clinical procedures on virtual simulation platforms. Each sequence is annotated with both global and step-level accuracy labels, compliance scores based on national clinical practice standards, detailed error classifications following a hierarchical taxonomy, and expert feedback text aligned with entry-level healthcare trainee cognitive levels. The construction of VMST represents a collaborative effort among computer science researchers, medical education specialists, and clinical practitioners, ensuring both technical rigor and pedagogical validity. Technical validation experiments demonstrate that models trained on VMST can effectively identify procedural errors and assess standards compliance, establishing a foundation for developing AI-assisted assessment tools that could transform medical simulation education in resource-constrained settings. Figure 1 illustrates the overall dataset construction workflow and the three-dimensional annotation framework.

Methods

Data Collection Environment

The VMST dataset was collected between September 2022 and June 2024 from three vocational medical colleges in eastern China, utilizing two commercial virtual simulation training platforms widely deployed across Chinese medical education institutions. Both platforms record detailed operation logs including action types, timestamps, spatial coordinates for relevant manipulations, and system-generated feedback events. Data collection was approved by the **Ethics Committee of Anhui Health College(Approval No. 2025013)**, and all participating students provided written informed consent for their anonymized training data to be used for research purposes.**All personal identifiers (names, student IDs, institution-specific information) were removed from the dataset prior to analysis. The dataset contains only operation sequences, timestamps, and performance metrics without any linkage to**

individual participants.All methods were carried out in accordance with relevant guidelines and regulations.

The simulation modules included in VMST were selected based on the National Health Industry Vocational Skills Standards, representing core competencies required for entry-level clinical practice. The 12 modules span four clinical domains: basic nursing procedures (venipuncture, urinary catheterization, nasogastric tube insertion, wound dressing), vital signs assessment (blood pressure measurement, pulse oximetry, ECG electrode placement, neurological examination), emergency response (cardiopulmonary resuscitation, airway management, hemorrhage control), and infection prevention (hand hygiene, sterile gowning, specimen collection). Table 2 presents the distribution of operation sequences across modules with corresponding annotation statistics, while Fig. 2 provides a visual representation of the data distribution across clinical domains and error rates.

Table 2
Distribution of simulation modules in the VMST dataset.

Module	Domain	Sequences	Mean Steps	Error Rate (%)
Venipuncture	Nursing	1,842	24.3	31.2
Urinary catheterization	Nursing	1,456	32.7	38.5
Nasogastric tube insertion	Nursing	1,287	28.1	42.1
Wound dressing	Nursing	1,523	19.6	25.8
Blood pressure measurement	Vital signs	1,634	15.2	22.4
ECG electrode placement	Vital signs	1,198	21.8	35.7
Cardiopulmonary resuscitation	Emergency	1,567	18.4	28.9
Airway management	Emergency	1,234	22.6	44.3
Hemorrhage control	Emergency	1,089	16.9	33.6
Hand hygiene	Infection prevention	1,345	12.4	18.7
Sterile gowning	Infection prevention	1,178	26.3	39.2
Specimen collection	Infection prevention	1,494	20.5	29.4
Total/Average	-	**15,847**	**21.6**	**32.5**

Error rate refers to the proportion of sequences containing at least one annotated procedural error.

Annotation Framework

The VMST annotation framework was developed through iterative consultation with a panel of eight clinical experts comprising three nursing education specialists with over 15 years of teaching

experience, two emergency medicine physicians, two infection control specialists, and one medical education researcher. The framework encompasses three complementary annotation dimensions designed to capture different aspects of procedural competence. The first dimension, Procedural Accuracy, assesses whether each action in a sequence is correctly executed in terms of technique, positioning, and timing. The second dimension, Scientific Compliance, evaluates the overall adherence of the procedure to established clinical practice guidelines and national standards. The third dimension, Error Classification, provides a structured taxonomy of common errors to support targeted feedback and remedial instruction.

Procedural accuracy annotations are assigned at the individual action level, with each action in a sequence labeled as Correct, Minor Error, Major Error, or Critical Error based on potential clinical consequences. Scientific compliance scores are computed using weighted checklists derived from national clinical skills examination standards, with each procedural step assigned a weight reflecting its clinical importance. Compliance scores range from 0 to 100. The error classification taxonomy comprises 6 primary categories (Sequence Errors, Technical Errors, Timing Errors, Safety Violations, Communication Lapses, and Documentation Errors), further subdivided into 24 secondary categories and 78 specific error types. Figure 3 illustrates the distribution of error types across the dataset.

Annotation Procedure and Quality Control

The annotation team consisted of 12 graduate students in nursing and clinical medicine who had completed at least two years of clinical internship, supplemented by 4 faculty members serving as senior reviewers. All annotators received a two-week training program. Each operation sequence was independently annotated by two trained annotators, with disagreements resolved through adjudication by a senior reviewer. Inter-annotator agreement was assessed using Cohen's kappa for categorical labels and intraclass correlation coefficient (ICC) for continuous scores. The final dataset achieved a mean Cohen's kappa of 0.83 (95% CI: 0.81–0.85) for action-level accuracy labels and an ICC of 0.91 (95% CI: 0.89–0.93) for compliance scores, indicating substantial to excellent agreement. The entire annotation process spanned approximately 18 months.

Data Records

The VMST dataset is publicly available through the Science Data Bank repository at https://doi.org/10.57760/sciencedb.36258. The repository contains all data files, documentation, and supplementary materials for research and development applications. Data are released under the Creative Commons Attribution 4.0 International license. Table 3 summarizes the file structure and contents of the dataset repository. Figure 4 presents the distribution of compliance scores across the dataset.

Table 3
Contents of the VMST dataset repository.

File/Directory	Format	Description
README.md	Markdown	Dataset overview, usage instructions, and citation information
vmst_sequences.json	JSON	Complete operation sequences with timestamped actions and metadata
vmst_annotations.json	JSON	Accuracy labels, compliance scores, and error classifications
error_taxonomy.csv	CSV	Hierarchical error classification taxonomy with definitions
train_val_test_split.json	JSON	Pre-defined data splits (70%/15%/15%)
scripts/	Python	Data loading utilities and baseline model implementations

Technical Validation

To evaluate the dataset's suitability for AI-based assessment, we implemented three baseline models: a Long Short-Term Memory (LSTM) network for sequence classification[20], a Transformer model adapted for operation sequence encoding[21], and a hybrid model combining convolutional feature extraction with recurrent sequence modeling. All models were trained using the pre-defined data splits (70% training, 15% validation, 15% test). Table 4 presents the performance of baseline models on the held-out test set, **and** Fig. 5 **provides a visual comparison of these metrics across different models**. The Transformer model achieved the best overall performance with 89.3% accuracy in action-level error detection and a mean absolute error of 4.7 points in compliance score prediction.

Table 4
Baseline model performance on the VMST test set.

Model	Action Acc. (%)	Action F1 (%)	Sequence Acc. (%)	Compliance MAE
LSTM	84.6	78.3	86.2	6.8
Transformer	89.3	84.7	91.7	4.7
CNN-LSTM	87.1	81.2	89.4	5.5
Random baseline	50.0	33.3	50.0	18.4

Acc. = Accuracy, MAE = Mean Absolute Error. All metrics computed on the held-out test set (n = 2,377 sequences).

Error analysis revealed systematic patterns informative for future model development. The most challenging error types to detect were timing-related errors and subtle technical mistakes requiring domain knowledge to identify. The Transformer model demonstrated particular strength in detecting

sequence errors. Cross-module analysis revealed that model performance was highest for modules with more proceduralized operations (hand hygiene, blood pressure measurement) and lowest for modules requiring complex judgment (airway management, nasogastric tube insertion). Figure 6 presents learning curves and cross-module performance analysis.

Usage Notes

The VMST dataset is intended for research and development of AI-assisted medical education assessment systems. Researchers may use this dataset to train models for automated error detection, compliance assessment, feedback generation, and learning analytics in virtual simulation environments. To ensure comparability across studies, model development should utilize the pre-defined train/validation/test splits. The data were collected from Chinese vocational medical students using specific simulation platforms, and models trained on VMST may require adaptation when deployed in other educational contexts.

Privacy and Ethical Considerations: All data in this dataset have been fully anonymized and contain no personal identifiers. The operation sequences and performance metrics cannot be traced back to individual students. Users of this dataset should maintain this privacy protection and refrain from any attempts to re-identify participants.

Discussion

The VMST dataset represents a significant contribution to the intersection of medical education and artificial intelligence, addressing a critical gap in annotated training data for AI-assisted procedural skill assessment. The demonstrated effectiveness of machine learning models trained on VMST points toward promising directions for developing intelligent tutoring systems that can provide real-time, personalized feedback to students practicing clinical procedures in virtual environments. The multi-dimensional annotation framework developed for VMST, encompassing accuracy, compliance, and error classification, provides a template that may be adapted for other domains where procedural skill acquisition is critical.

In conclusion, the VMST dataset provides a foundational resource for developing AI systems capable of assessing procedural accuracy and scientific compliance in virtual medical simulation training. The rigorous annotation procedures, comprehensive error taxonomy, and demonstrated utility for machine learning model development position VMST as a valuable tool for researchers and developers working at the intersection of artificial intelligence and medical education.

Declarations

Competing Interests

The authors declare no competing interests.

Author Contribution

CY Wang. Conceptualization, Methodology, Data Curation, Writing - Original Draft. Software, Validation, Formal Analysis. Li Zhang. Investigation, Resources, Writing - Review & Editing. Supervision, Project Administration, Funding Acquisition. All authors reviewed and approved the final manuscript.

Acknowledgement

This work was supported by the Key Program of Natural Science Research of Anhui Provincial Department of Education (Grant No. KJ2021A1564), the Provincial Quality Engineering Project – Physiology Teaching Innovation Team (Grant No. 2022cxtd059) and the Key Project for the Cultivation of Excellent Young Teachers by Anhui Provincial Department of Education (Grant No. YQZD2024099).

Data Availability

The VMST dataset described in this paper is publicly available through the Science Data Bank repository at https://doi.org/10.57760/sciencedb.36258. Data are released under the Creative Commons Attribution 4.0 International license.

Code Availability

Python code for data loading, preprocessing, and baseline model implementations is available at https://github.com /wangchaoya3/VMST.

References

1. World Health Organization. Global strategy on human resources for health: Workforce 2030. Geneva: World Health Organization; 2016.
2. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010; 376(9756): 1923–1958.
3. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011; 86(6): 706–711.
4. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011; 306(9): 978–988.
5. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005; 27(1): 10–28.

6. Barsuk JH, McGaghie WC, Cohen ER, O'Leary KJ, Wayne DB. Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. Crit Care Med 2009; 37(10): 2697–2701.
7. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB.A critical review of simulation-based mastery learning with translational outcomes.Medical Education, 2014; 48(4): 375–385.
8. Sandars J, Correia R, Dankbaar M, et al. Twelve tips for rapidly migrating to online learning during the COVID-19 pandemic. MedEdPublish 2020; 9: 82.
9. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education: 2003–2009. Med Educ 2010; 44(1): 50–63.
10. Cant RP, Cooper SJ. Use of simulation-based learning in undergraduate nurse education: An umbrella systematic review. Nurse Educ Today 2017; 49: 63–71.
11. World Federation for Medical Education. Basic Medical Education: WFME Global Standards for Quality Improvement. Copenhagen: WFME; 2020.
12. Chen X, Xie H, Zou D, Hwang GJ. Application and theory gaps during the rise of artificial intelligence in education. Comput Educ Artif Intell 2020; 1: 100002.
13. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv preprint arXiv:2009.13081 (2020).
14. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A dataset for biomedical research question answering. In: Proceedings of EMNLP-IJCNLP 2019: 2567–2577.
15. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of CHIL 2022: 248–260.
16. Kuzilek J, Hlosta M, Zdrahal Z. Open University Learning Analytics dataset. Sci Data 2017; 4: 170171.
17. Welbl J, Liu NF, Gardner M. Crowdsourcing Multiple Choice Science Questions. In: Proceedings of the 3rd Workshop on Noisy User-generated Text (W-NUT 2017): 94–106.
18. Downing SM, Yudkowsky R (eds.). Assessment in Health Professions Education (2nd Edition). New York: Routledge; 2019.
19. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS One 2020; 15(2): e0229596.
20. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9(8): 1735–1780.
21. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 2017; 30: 5998–6008.

Figures

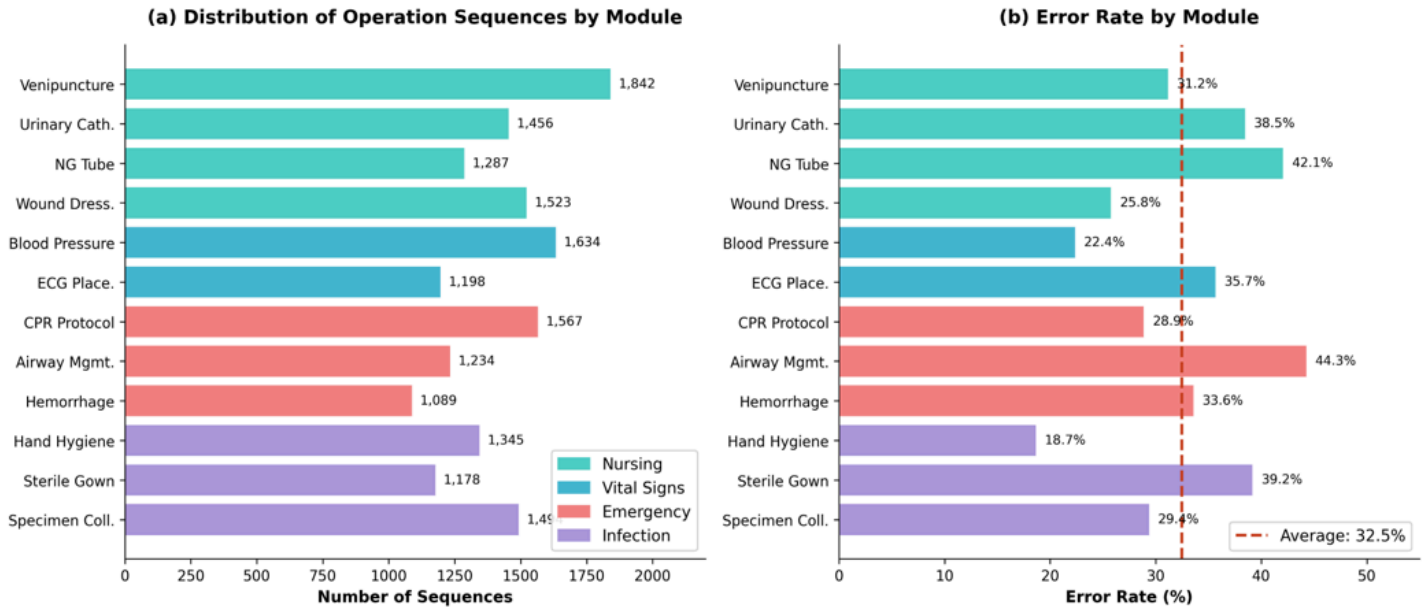


Figure 1

Distribution of operation sequences across clinical domains and corresponding error rates by simulation module.

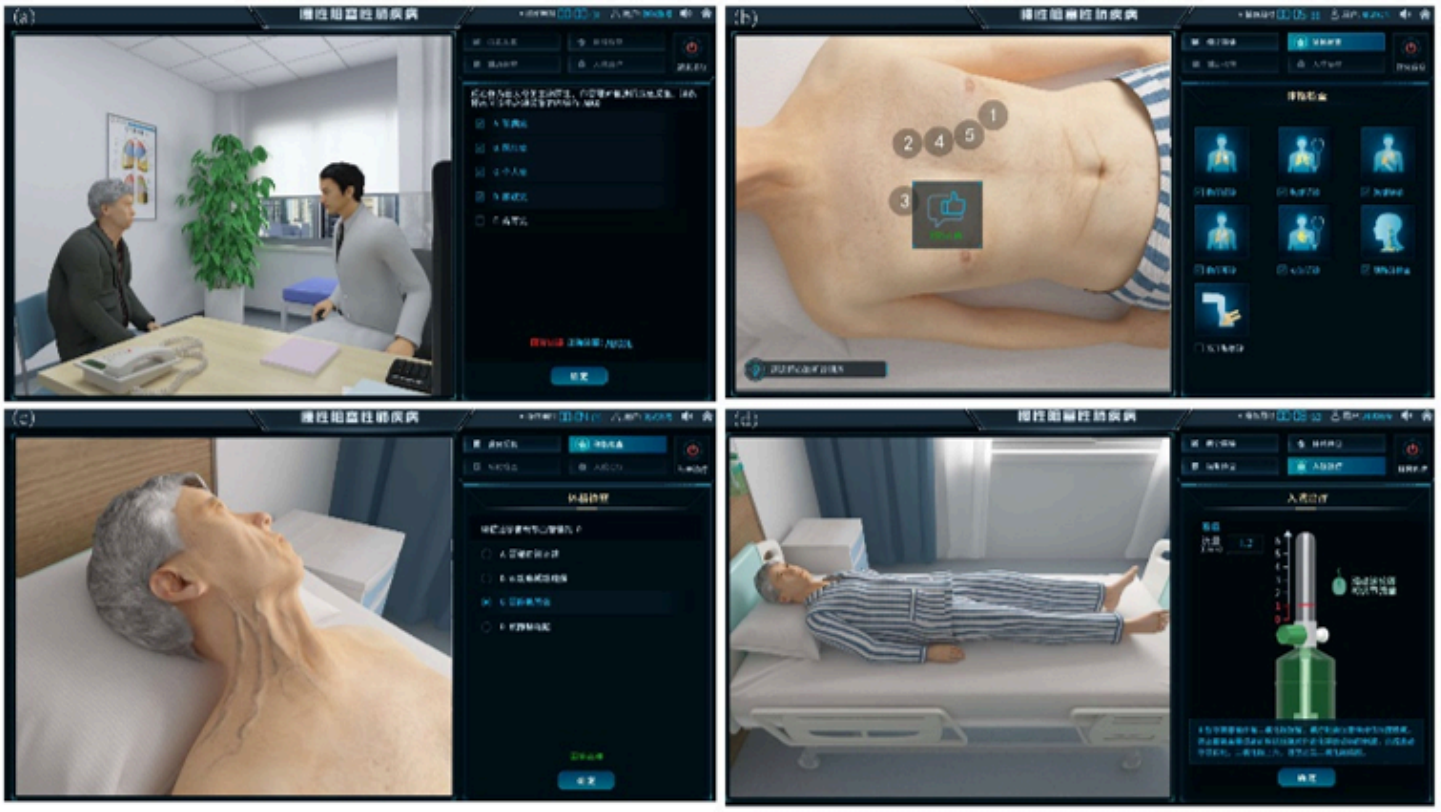


Figure 2

Virtual medical simulation training platform interface (COPD module example)

(a) History taking: Students interact with virtual patients as attending physicians; the right panel displays a medical history checklist with real-time feedback on selections. (b) Physical examination: Anatomical thorax view with five standardized auscultation points; students complete examinations in correct sequence. (c) Clinical reasoning: Students answer clinical judgment questions based on examination findings with immediate accuracy feedback. (d) Treatment implementation: Oxygen therapy interface where students adjust flow rates according to Type II respiratory failure treatment protocols. Operation duration is displayed in real-time, and all operational data are automatically recorded for compliance assessment.

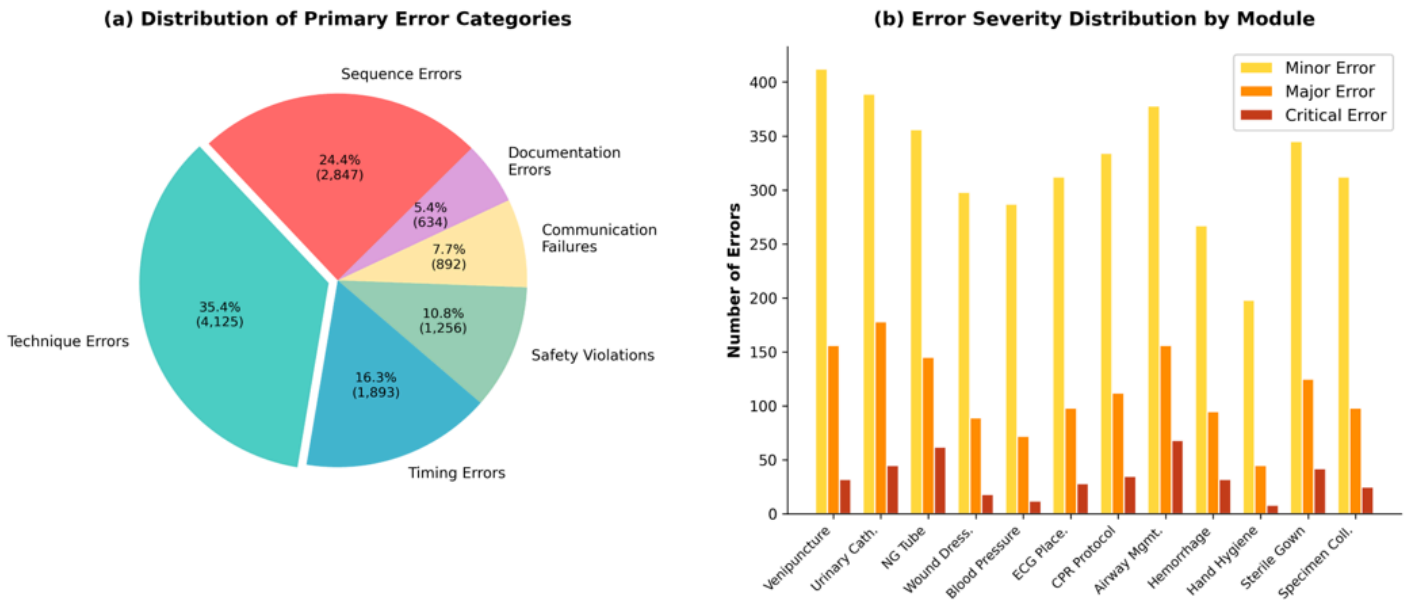


Figure 3

Distribution of error categories and severity levels across simulation modules.

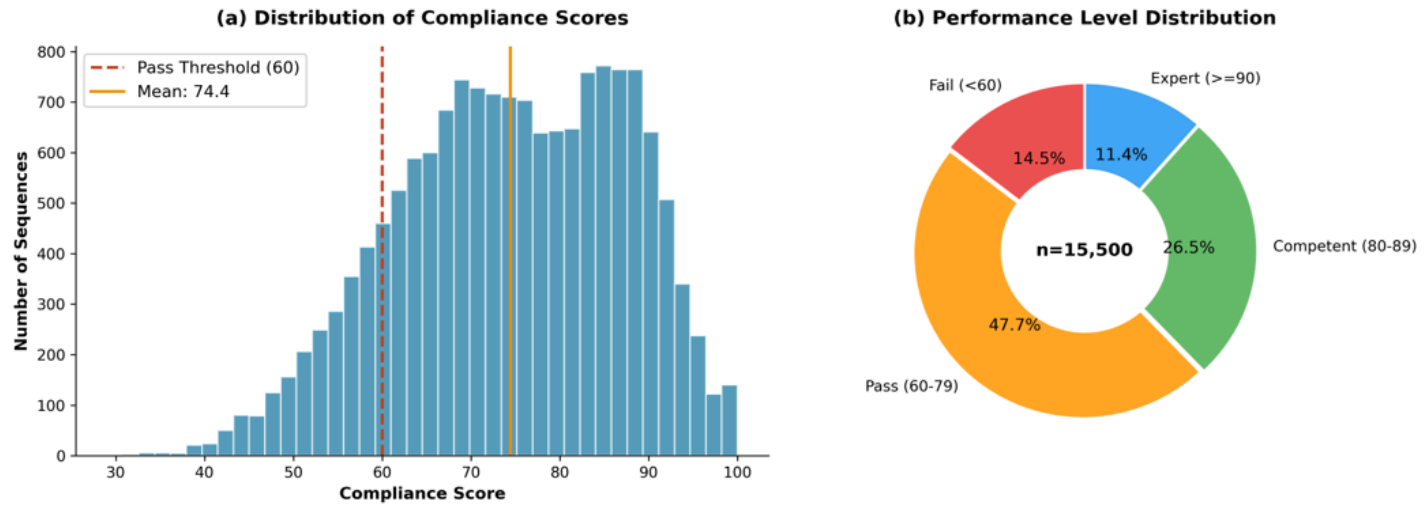


Figure 4

Distribution of compliance scores and performance grades in the VMST dataset.

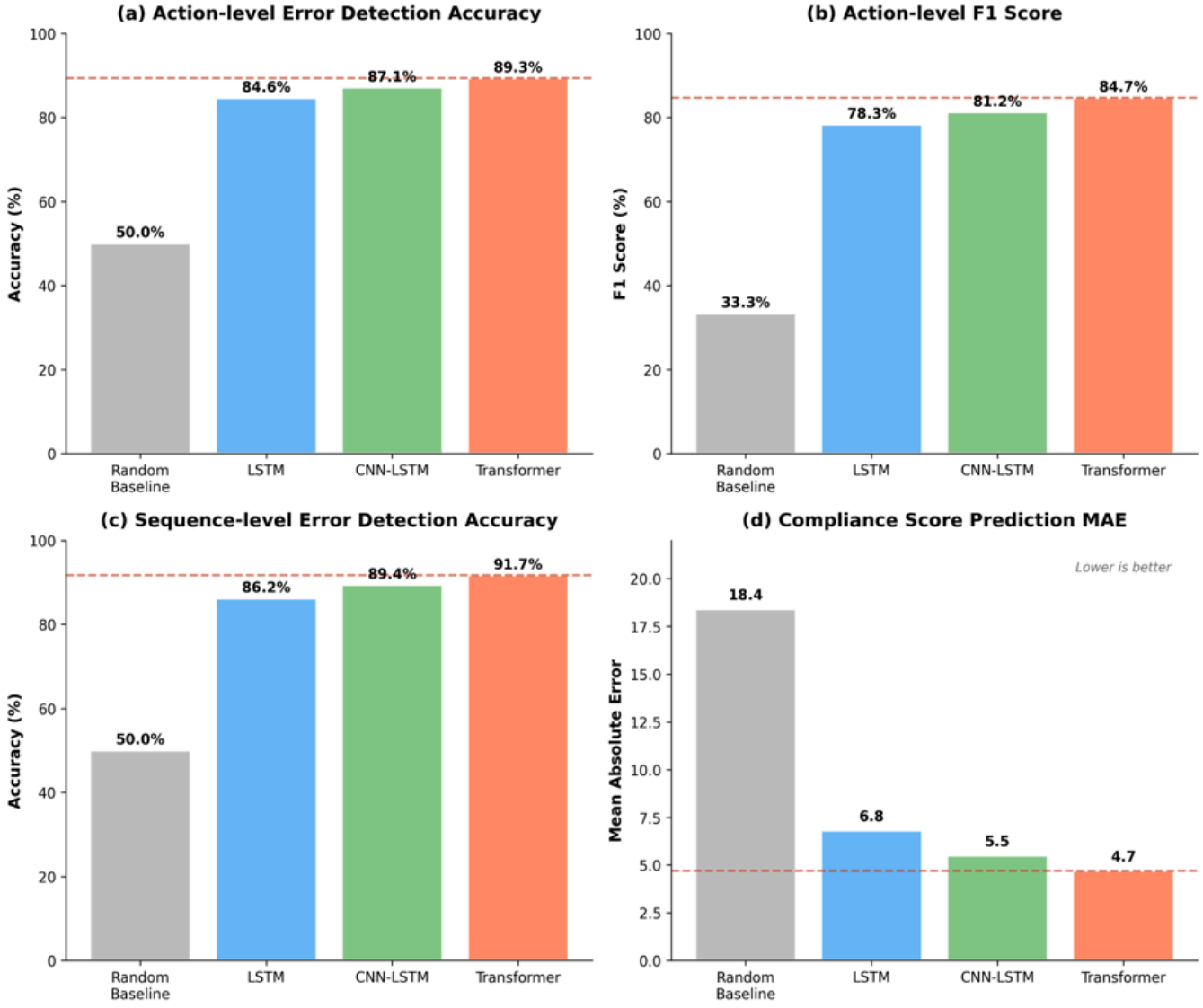


Figure 5

Performance comparison of baseline models across evaluation metrics.

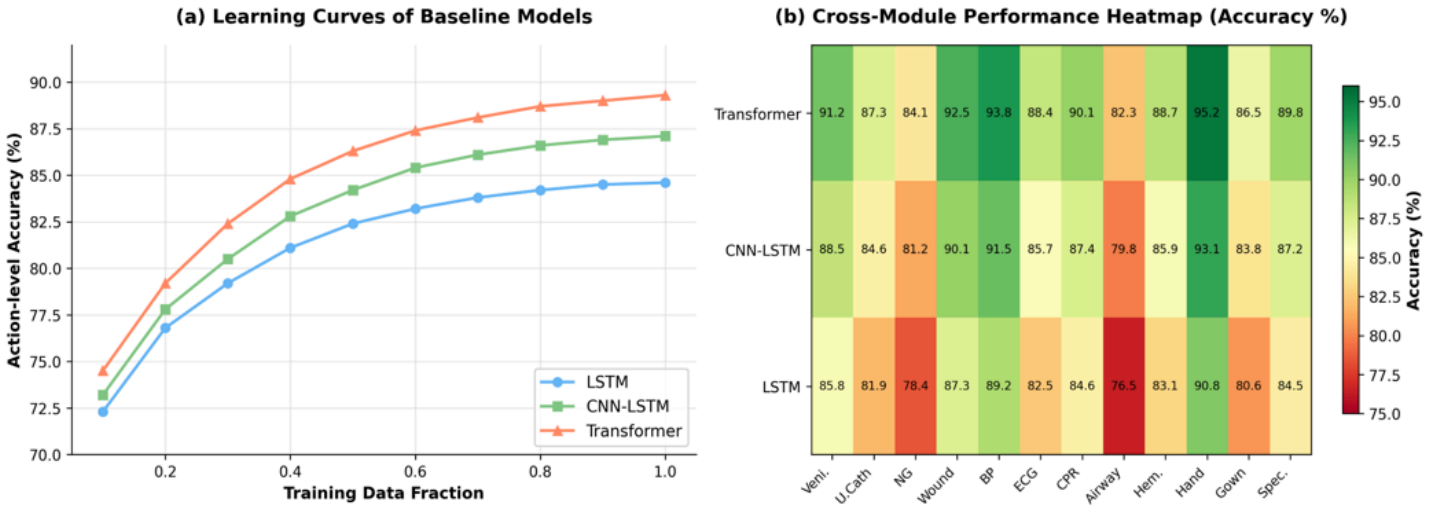


Figure 6

Learning curves for baseline models and cross-module performance heatmap.