INTERNATIONAL JOURNAL OF EDUCATIONAL REVIEW
Vol.1, Issue 3 (2025)
ISSN: 2760-2427(Print) 2760-2435 (Online)
HOMEPAGE: https://www.janfs.org

Design and Preliminary Application of a CV-Based Multimodal Teaching Support System for Higher Vocational Education

Yang Xiaoxue
Wuhan Vocational College of Software and Engineering, Wuhan, Hubei, 430205
Email: shirly520123@gmail.com

KEYWORDS

Computer Vision; Multimodal Educational Technology; Higher Vocational Education; Teaching Assistance; Intelligent Classroom

ARTICLE HISTORY
Received: 09/04/2025
Revised: 09/13/2025
Accepted: 09/15/2025

ABSTRACT

In recent years, the steady progress of artificial intelligence has brought computer vision (CV) into the spotlight of educational research. Compared with traditional approaches that mainly depend on text or speech, CV is capable of processing multiple information streams—such as images, recognized text, and structural features—and reorganizing them into meaningful teaching resources. This ability makes it possible to support classroom instruction in a way that is more visual, interactive, and efficient. For higher vocational education, where learning tasks are strongly practice-oriented and course materials often contain diagrams or graphical elements, such technologies are particularly relevant. This paper examines how CV-based multimodal techniques can be applied in vocational teaching, with specific attention to resource management, knowledge extraction, and interactive support in the classroom. A prototype framework was designed, integrating functions such as network topology identification, keyword extraction from slides, and the digital capture of handwriting from blackboard work. When tested in teaching scenarios, the system showed promise in improving students' understanding and participation, while also reducing repetitive tasks for instructors. The study therefore provides both a conceptual foundation and preliminary practical evidence for advancing the digital transformation of vocational education.

1. Introduction

1.1 The Need for Digital Transformation in Higher Vocational Education

The rapid evolution of information and communication technologies is reshaping the landscape of higher vocational education, gradually moving it away from

Copyright 2025 by author(s) This work is license under the CC BY 4.0
DOI:10.64583/b0j1xn67

conventional classroom practices and toward digital, intelligent, and more interactive modes of learning. Vocational training, by nature, stresses hands-on practice and direct application, which creates higher expectations for timely feedback, abundant resources, and immersive learning experiences. In recent years, government initiatives have also underscored the importance of “digital empowerment” and “integration between industry and education,” further accelerating the incorporation of emerging technologies into daily teaching. Against this backdrop, a key question emerges: how can new technologies be effectively employed not only to visualize complex instructional content but also to stimulate active learning and improve the overall student experience? Addressing this question has become central to advancing both the quality and sustainability of vocational education.

1.2 Gaps in Traditional Teaching and Challenges in Practice-Oriented Courses

Although the adoption of multimedia tools and online platforms has improved classroom delivery to some extent, substantial limitations remain in practice-oriented courses(Huang, Li, & Zheng, 2025). Instructional materials are still frequently confined to static slides and textbook illustrations. These resources lack intelligent processing capabilities, leaving teachers with few tools to highlight essential concepts in real time. Moreover, many abstract topics—such as computer network topologies, routing mechanisms, or communication protocols—are highly dependent on visual representations, yet students often struggle to construct accurate mental models from simple diagrams or text-heavy explanations. Another persistent problem is the transient nature of classroom activities: blackboard notes and in-class demonstrations are rarely recorded or archived, depriving students of valuable review materials. Such deficiencies reduce the practical effectiveness of vocational training, weaken students’ ability to consolidate knowledge, and ultimately hinder the transfer of classroom learning into real-world professional skills.

1.3 Research Purpose and Novelty

To address these challenges, the present study investigates the application of computer vision (CV)-based multimodal technologies in vocational education. The primary objectives are threefold:

To design a CV-supported framework capable of automatically identifying, extracting, and reorganizing instructional resources in ways that better suit the demands of vocational courses; To develop and evaluate functional modules—including topology recognition, slide-based keyword extraction, and blackboard content digitization—that aim to improve the clarity, accessibility, and informatization of classroom teaching; To test the effectiveness of these modules in authentic teaching contexts, with particular attention to student participation, learning outcomes, and the reduction of teachers’ repetitive workloads.

The study makes several contributions. First, it introduces multimodal CV techniques into vocational education in a structured and systematic manner, extending the scope of educational technology beyond traditional text- and speech-based approaches. Second, it proposes practical strategies for visual resource recognition and processing that directly respond to the graphical and experimental characteristics of vocational curricula (Yang, Bu, & Li, 2025). Finally, the research establishes a meaningful link between technological development and classroom practice, laying a foundation for future teaching-assistance models that can be adapted, scaled, and continuously optimized across diverse institutions and disciplines.

2. Literature Review

2.1 Applications of Computer Vision in Educational Technology

During the last decade, computer vision (CV) has gradually moved from technical laboratories into educational settings. Researchers abroad have experimented with diverse applications, ranging from automated grading to behavioral analysis. For instance, optical mark recognition is now widely used to score multiple-choice exam sheets, a practice that has saved instructors countless hours. Other projects have employed video analytics to map students' attention shifts or even track patterns of collaboration within groups. In the context of laboratory safety, CV-based detection systems have been piloted to identify whether students are wearing protective gear or following required procedures.

China has also witnessed a growing interest in CV for education, though the initiatives tend to remain exploratory. Some teams have tested blackboard recognition systems; others have looked at classroom expression monitoring to gauge student engagement. Even attendance-taking through facial recognition has been reported in several institutions. These studies illustrate the flexibility of CV across different environments. Yet, as many practitioners point out, most of these applications feel more like isolated experiments than integrated solutions. Teachers often find them useful for demonstration but not easy to incorporate into their daily routines. This mismatch reveals that while CV is promising, its educational role is still developing and somewhat fragmented.

2.2 The Value of Multimodal Information Processing in Teaching

Classroom teaching is never confined to a single mode of communication. A teacher may speak, write on the board, gesture toward a diagram, and respond to students' questions—all within a few minutes. Because of this complexity, multimodal processing has become a vital research direction. Integrating images with text or combining transcripts with speech recordings can create resources that go beyond what any single channel provides (Wang, Wang, & Zheng, 2021). Automatic lecture summaries and knowledge navigation maps are two examples that have already shown practical value.

However, CV-dominant multimodality remains underexplored. Unlike approaches that treat text or audio as the main channel, CV-oriented methods take images and diagrams as the central input, adding OCR text or structural features as supportive layers. This could be particularly powerful in visually intensive courses. Even so, many of the systems developed so far stop at the technical demonstration stage. A common frustration expressed by teachers is that while prototypes may generate neat outputs, they rarely align with actual classroom needs. Some blackboard recognition models, for example, struggle with messy handwriting or diagrams drawn in haste, which reduces their reliability. In other words, the technology often looks polished in conference papers but far less convincing in real classrooms.

2.3 Characteristics and Challenges of Higher Vocational Education Curricula

Higher vocational education is distinct from general academic programs because it focuses on cultivating job-ready skills. Students are not only expected to understand theories but also to demonstrate competence in authentic practice scenarios. This dual demand creates both opportunities for innovation and challenges in implementation.

On the one hand, vocational curricula are deeply practice-oriented. Teachers frequently rely on demonstrations, case studies, and repeated training exercises. Yet, static slides and textbook figures cannot capture the dynamism of real processes. As a result, many students resort to rote memorization, which undermines the very goal of skill integration.

On the other hand, abstract concepts pose particular difficulties. Networking courses are a clear example: topological structures, routing mechanisms, and data flows can be extremely hard to visualize. Blackboard sketches and textual explanations are rarely sufficient, and even complex diagrams risk overwhelming students with detail. Teachers often report that learners either “get lost in the arrows and boxes” or simply memorize without real comprehension.

Finally, maintaining student motivation is a continuous challenge. Many vocational learners rely heavily on direct visual input and timely feedback to stay engaged. Traditional lecture-based models often fail to provide this, leading to passive behaviors such as “watching without doing” or “doing without understanding.” Such patterns are especially damaging in vocational contexts, where active participation is crucial. Together, these issues reveal the pressing need for technological support that can make abstract concepts visible, preserve teaching traces, and keep students actively involved.

2.4 Research Gaps and Points of Entry

Despite increasing attention, the current body of literature leaves several gaps when viewed through the lens of vocational education. Much of the work has focused on primary, secondary, or general university settings, where exam grading or classroom monitoring dominate. The specific requirements of vocational curricula—skills training, operational competence, and hands-on integration—are less frequently addressed.

At the application level, most studies deal with single, isolated tasks. A system might perform blackboard recognition, another might analyze student expressions, but these remain siloed. What vocational teaching requires is not a collection of isolated tools but a connected framework, one that links topology recognition with slide-based keyword extraction and the archiving of lecture traces. Only when these modules function together can technology offer real classroom value.

Another weakness is the shortage of long-term, empirical validation. Many published studies showcase impressive prototypes, yet they rarely track actual improvements in student comprehension or reductions in teachers’ workload.

Questions that matter most to educators—whether these systems genuinely make abstract topics clearer, or whether they help save preparation time—are often left unaddressed. Some teachers even report that trial systems increase their burden, as they must correct recognition errors or manage additional software.

To respond to these issues, this study selects information technology courses in vocational education as a test case. The proposed CV-based multimodal framework is deliberately designed to match the graphical and practice-oriented character of such courses. By integrating multiple modules into a coherent system and testing it with experimental and control groups, the research aims not only to extend methodological innovation but also to provide practical evidence of effectiveness. In doing so, it attempts to bridge the gap between technological promise and classroom reality, a step that is often missing in current literature.

3. Theoretical Framework and Research Approach

3.1 Definition and Scope of Multimodal Educational Technology

In a typical networking class, the same concept often appears in three guises: a rough star-topology sketch on the board, a polished diagram on the following slide, and a passing remark that "links are swapped at the hub." Multimodal educational technology, as used in this study, refers to the attempt to stitch together such scattered traces—visual, textual, and behavioral—into reusable learning units.

While in principle multimodality could involve speech, gesture, or even biometric data, our focus is on computer vision (CV)-driven fusion, because vocational courses are unusually visual. We prioritize three modalities: (i) the image stream, covering slides, topology diagrams, and lab photos; (ii) the text stream, extracted from board notes or slides via OCR; and (iii) structured features, such as classification results, keyword lists, or semantic tags. Once integrated, these resources no longer remain fragmented. Instead, they are reorganized into "review cards" or linked knowledge nodes that students can revisit after class.

It is worth noting that this definition is not merely aspirational. We explicitly set aside other multimodal channels such as speech recognition, because pilot runs showed that noisy labs and heavy accents produced unstable transcripts that required more editing than they saved. OCR and lightweight image analysis, in contrast, proved robust and easy for instructors to tolerate (Wang & Xu, 2024). This kind of pragmatic narrowing of scope is essential if multimodality is to serve education rather than remain a technical showcase.

3.2 Pedagogical Anchors for the Study

Educational theory is not included here for decoration; it acts as a day-to-day guardrail against technological drift. Three perspectives shape the present framework.

First, the TPACK model (Technological Pedagogical Content Knowledge) reminds us that CV tools have to sit within a triangle of subject content and teaching method. In our setting, this means aligning topology recognition with the actual learning objectives of network fundamentals, and pairing it with pedagogical strategies such as case demonstrations or problem-based lab tasks.

Second, Cognitive Load Theory alerts us to the hidden cost of excessive information. Students in networking classes often get lost when diagrams exceed a certain size—roughly more than eight labeled nodes or three layers of arrows. Our system therefore applies a rule of thumb: complex diagrams are broken down into progressive reveals (first functions, then links, then flows). In early pilot sessions, this cut the number of clarification questions about diagrams almost in half.

Third, the idea of human–machine collaboration shapes how the system is positioned in class. Teachers focus on logic, pacing, and interaction, while the system performs repetitive tasks: capturing notes, generating keyword prompts, archiving diagrams. Importantly, the system does not overwrite a teacher's emphasis. If OCR confidence drops below 0.85 or a symbol is ambiguous, the item is stored as a clipped image with a timecode, leaving the decision to the instructor. This safeguard prevents the system from dictating the lesson and reassures teachers that they remain in control.

3.3 A Working Integration Scheme for CV in Vocational Education

Rather than a neat four-stage pipeline, our integration behaves more like a gated weave. Tasks do not always flow forward smoothly; some are paused, flagged, or routed differently depending on classroom conditions.

- Input. The system ingests whatever is available: slides, ad hoc board sketches, photos of lab setups, even snapshots taken by teaching assistants. Lighting and glare are variable, so input quality is uneven.
- Processing. Topology recognition runs first. If shapes are borderline (e.g., star vs. snowflake), the classifier attaches a "verify" tag and skips forced labeling. OCR then parses board notes; items below 0.85 confidence are preserved as timestamped image clips rather than text. Only after this gated pass do we fuse items into structured "review cards."
- Application. In practice, when students confuse ring and bus topologies, the teacher can tap a card; the system overlays the corresponding diagram, highlights the missed link rule, and reconnects it to the earlier slide. After class, those cards are bundled into short recap sets (≤10 items) rather than long transcripts, which students report are easier to study from.
- Outcome. The expected benefit is fewer repeat explanations and more concise review material. But caveats remain: dim lighting and glossy boards reduce OCR accuracy; some teachers prefer handwriting over slides, which generates noisy inputs. To address this, we distributed dark-background board templates and built a one-click capture tool for assistants.

Finally, the integration is framed by data and governance rules. All captured artifacts follow an opt-in consent protocol: raw images are retained for 14 days by default, after which they are deleted; only derived text and annotated diagrams persist as course metadata. This explicit boundary-setting is crucial for teacher acceptance, and is often overlooked in purely technical accounts of "intelligent classrooms."

4. Research Strategy and Methodology

4.1 Topology Recognition Module

Network topology is a cornerstone concept in computer networking but also one of the most abstract. Instructors often complain that students mix up ring, bus, and star structures even after repeated explanations. To ease this, we built a topology recognition module that processes both slides and board sketches.

For the backbone, we compared several CNN models. VGG16 was quickly discarded: on our classroom machine (a single GTX 1060), inference took ~2.5 seconds per image, which was unacceptable in live lectures. ResNet18 achieved ~84% accuracy on a hand-drawn dataset of 600 student sketches, with inference under 0.5 seconds. MobileNetV2 was slightly faster (0.3s) but slightly less accurate (~82%). We chose ResNet18 as the default, with MobileNetV2 as fallback on lower-end devices.

But accuracy numbers only tell part of the story. In practice, students' sketches are messy. A star missing one edge was often predicted as a tree. Rather than force labels, we built a "verify" flag: if confidence <0.80, the system outputs the image as-is and prompts the teacher (He & Mi, 2025). During one pilot, a snowflake-like sketch confused both the model and half the class—our overlay highlighted the missing hub, turning the mistake into a discussion point. Teachers later told us they appreciated that the system "admitted uncertainty" instead of misleading everyone.

4.2 PPT Keyword Extraction Module

Vocational slides tend to be overloaded: multiple protocols, acronyms, and diagrams squeezed into a single page. Students try to copy everything and miss the key points. Our keyword extraction module combines OCR with TF–IDF ranking to pull out ~5–10 keywords per slide.We initially tested neural abstractive summarizers but abandoned them: bilingual slides (English technical terms + Chinese notes) confused the models, and generation lag (~3–4s) broke the lecture flow. OCR + lightweight ranking produced predictable results in <1s. Importantly, teachers can override the system: they can "pin" important terms (e.g., TCP/IP) or discard irrelevant ones (e.g., random slide footers). This manual control reassured instructors that the system would not dictate emphasis.

Feedback was mixed but useful. Many students said the real-time keywords helped them reorient when they lost track. One class reported fewer "what did you just say?" interruptions. But some complained that keywords alone lacked context—"routing table" appeared without explanation, leaving them uncertain. Teachers also raised a concern: the overlay of keywords on slides sometimes distracted from diagrams. In response, we added an option to export keywords into a recap sheet instead of projecting them live.

4.3 Blackboard Recognition Module

Blackboard writing remains indispensable in vocational classes, especially for step-by-step derivations. Yet chalk notes disappear once erased. Our blackboard recognition module captures these traces through mounted cameras.

Preprocessing (denoising, contrast adjustment, edge detection) is followed by handwriting OCR. Results vary wildly: chalk dust, cursive writing, and arrows overlapping text often reduce accuracy. In one trial under fluorescent glare, accuracy dropped to ~60%, far below usable. We responded with dynamic thresholding and a fallback: tokens below 0.85 confidence are saved as cropped images with timecodes.

Students valued the resulting “knowledge traces.” Being able to replay how a formula was derived—rather than just the final result—was cited as “a lifesaver” in exam prep(Jiang & Sun, 2025). Teachers also reused digitized notes the following semester. But problems remain: one instructor joked that her chalk writing was misread as three different English words; another worried that constant camera capture could intimidate shy students. These comments remind us that technical feasibility does not guarantee pedagogical acceptance.

4.4 From Data Flow to Classroom Use

Rather than the rigid “four layers” often depicted in system diagrams, our framework behaves more like a weave with checkpoints.

- Gathering Inputs. The system accepts whatever is available: slides, sketches, snapshots of lab setups—even phone photos submitted by students. Quality is inconsistent, so metadata such as device type and resolution are stored.
- Making Sense of Noisy Data. Recognition runs in stages. Topologies are classified first; uncertain cases are tagged rather than mislabeled. OCR follows, with low-confidence text preserved as images. Fusion then produces compact “review cards.”
- In-Class Use. Teachers can project overlays when misconceptions arise, or ignore them if flow matters more. Students receive recap sets capped at ~10 items/session; one group said they were “less overwhelming than a transcript.” Yet not everyone was happy: a few learners asked for longer summaries, suggesting the need for personalization.
- Closing the Loop. Logs capture how often teachers accept suggested keywords, or how many students revisit review cards after class. These are not just for algorithm tuning; they inform pedagogy. For example, frequent replays of subnet-mask diagrams flagged a gap in prior knowledge.

4.5 Data Governance and Practical Constraints

Technology alone does not decide adoption—policies and perceptions do. In our pilot college, the IT department insisted that all raw video remain on local servers, not cloud platforms. Teachers also requested signed student consent before recording, which delayed deployment by two weeks. As a compromise, we set retention to 14 days for raw images, while derived text and diagrams persisted as course metadata.Interestingly, teachers accepted the system more readily once governance was explicit. One put it bluntly: “I don’t want to discover next year that my blackboard doodles are on YouTube.” Such concerns show that intelligent classrooms require not just algorithms but social contracts.

5. Application Scenarios and Experimental Design

5.1 Experimental Context: Information Technology Courses

We chose Fundamentals of Computer Networking as the pilot course. Each class had about 48 students, and the facilities were modest: two labs with 25 working PCs each, some quite outdated. Teachers frequently noted that students struggled with the "invisible" nature of networks—data packets, routing paths, or topological differences. These were ideal pain points for testing whether CV-based assistance could help.

Networking was not the only candidate. Operating systems and database fundamentals showed similar needs, but networking stood out because diagrams, flows, and protocols dominate its instruction. The trial spanned a 16-week semester, covering 12 units. Some classes were recorded in labs; others were standard lecture halls with poor lighting, which created very different conditions for the system.

5.2 Teacher-Side Functions in Practice

Teachers had mixed feelings. Several enjoyed not having to redraw topologies—"finally I can save my chalk for real explanations," one said. But others disliked the timing: overlays sometimes popped up just as they were building suspense for a question, which broke their rhythm. One senior instructor even disabled the overlay function halfway through the semester, preferring to rely on recap sheets instead.

The auto-generated "course summaries" were convenient but also raised concern. A few teachers worried that students had stopped taking their own notes: "I looked around and saw half the class staring at the system's recap screen instead of writing anything." While time-saving, the summaries shifted classroom habits in unexpected ways.

5.3 Student-Side Experience

Students valued the recap sets most. Getting 8–10 cards per lecture was less overwhelming than a transcript. One group joked that they "revised on the bus as if they were flipping Pokémon cards." But not everyone was convinced: a few deliberately ignored the cards, saying they trusted their handwritten notes more.

Real-time keywords also split opinions. Many used them to stay on track, but several students found them distracting when they appeared over diagrams(Huang & Liang, 2025). One day, the projector lagged by three seconds, causing students to laugh and the teacher to shut the feature off in frustration. In labs, however, the digitized diagrams were praised; groups could compare their wiring against the overlay, catching mistakes earlier. Still, some students felt uneasy about being filmed all the time, and one remarked, half-jokingly, "It feels like Big Brother is also enrolled in our class."

5.4 Comparative Design and Evaluation

Two parallel classes served as experimental and control groups, both around 45 students, average age 19. The experimental class used the CV system throughout; the control class followed conventional methods.

Results were uneven. On final exams, the experimental group's average was about 8% higher, but variation was large: in some units the gap was 15%, in others negligible. Lab tasks showed clearer benefits: average completion times were shorter, but three groups still struggled with basic cabling despite having overlays. Engagement logs revealed heavy use of recap cards (over 70% of students accessed them weekly), but classroom discussions did not increase much. Teachers saved time on redrawing diagrams but spent extra time troubleshooting OCR misreads.

In short, the system improved clarity and saved effort in some areas, but it did not magically transform participation. Adoption turned out to be situational—dependent on teacher style, classroom conditions, and even projector performance.

6. Preliminary Conclusions and Future Prospects

6.1 Research Conclusions

This study carried out a preliminary but systematic exploration of how computer vision (CV)-based multimodal technologies may support higher vocational education. After surveying both domestic and international research, we proposed a CV-centered framework tailored to the practical features of vocational curricula. Key modules—such as topology recognition, slide keyword extraction, and blackboard digitization—proved technically feasible in pilot testing.

Beyond feasibility, these modules demonstrated clear pedagogical value. Students found it easier to grasp abstract concepts when visual overlays and recap sets were available, and teachers reported less repetitive effort in redrawing diagrams or reorganizing board notes. Compared with prior intelligent education research that mainly emphasized text or speech modalities, this work highlighted the visual modality as the central driver—arguably more suitable for practice-oriented and graphically rich courses. Overall, the study indicates that CV-based multimodal assistance has both theoretical significance and practical utility, laying a foundation for future scaling.

6.2 Research Limitations

Several limitations need to be acknowledged in interpreting these findings. On the data side, the models were trained on relatively small and self-constructed datasets—for example, only about 600 student sketches were available for topology recognition—which inevitably raises concerns about generalization. In terms of scope, the present study was conducted within the domain of information technology courses, and the extent to which the framework can be applied to other fields such as nursing, mechanical engineering, or applied arts remains untested. Another limitation lies in the duration of the pilot: with only one semester of implementation, it is difficult to draw conclusions about long-term effects on sustained learning outcomes, skill transfer, or changes in teaching workload.

In addition to these technical and methodological constraints, there are also social and cultural considerations. Teachers and students did not respond uniformly; some embraced the system with enthusiasm, while others expressed hesitation. Concerns ranged from over-reliance on system-generated summaries, which might weaken independent note-taking, to unease about the constant presence of classroom cameras. These reactions suggest that future scaling of the system will depend not only on algorithmic improvements but also on thoughtful management of acceptance and trust.

6.3 Future Directions

Future research should build on these lessons in several ways. One urgent task is the development of large-scale, multi-disciplinary, and multi-context datasets for vocational education. Such resources would significantly improve model robustness and enable replication across different institutional settings. Another important avenue is the expansion of application scope. While this study focused on information technology, vocational courses in areas such as mechanical engineering, electrical engineering, or nursing contain distinctive visual resources—ranging from circuit diagrams to anatomical sketches—that can test and extend the adaptability of CV-based approaches.

Equally critical is the integration of CV technologies with large language models. Combining visual recognition with generative language capabilities could enable new functions, such as automatic drafting of lab reports, the creation of structured learning summaries, or personalized feedback for students(Zhang et al., 2023). At the same time, this direction requires safeguards against over-dependence, ensuring that learners continue to engage actively with core concepts rather than simply consuming machine-generated content.

In parallel, evaluation and deployment strategies need to evolve. Short-term pilots must give way to long-term, real-classroom studies that establish a closed loop from data collection to intelligent processing, instructional support, and feedback optimization. Cost–benefit analysis, infrastructure requirements, and teacher training should be considered alongside algorithmic refinement, since sustainability hinges on practical integration into everyday teaching. Finally, issues of governance and ethics must not be overlooked. Questions about data storage, retention policies, consent mechanisms, and teacher autonomy are central to whether intelligent classrooms will gain acceptance. Some partner institutions have even suggested involving student–teacher committees in co-designing usage protocols, which could serve as a model for building legitimacy and trust.

6.4 Concluding Remarks

In summary, CV-based multimodal technologies offer a promising pathway for the modernization of vocational education. The study demonstrates that they can improve conceptual clarity, preserve valuable instructional traces, and reduce repetitive workload, while also surfacing challenges related to adoption and governance. These technologies should be seen not as replacements for existing pedagogy but as strategic complements that support teachers and empower students.

With continued data enrichment, cross-disciplinary trials, and careful integration of governance measures, this line of research is expected to show wider potential across vocational domains. The ultimate goal is to guide higher vocational education toward more intelligent, responsive, and learner-centered development—one that respects existing practices while opening new opportunities for digital transformation.

References

He, M., & Mi, H. (2025). Advantages, concerns, and prospects of applying multimodal large models to ideological and political education. School Party Building and Ideological Education, (11), 79–82. https://doi.org/10.19865/j.cnki.xxdj.2025.11.017

Huang, W., & Liang, G. (2025). Intelligent construction of multimodal ethnic mathematics education resources and curriculum practice. Journal of Primitive Ethnic Culture, 17(4), 143–152. https://doi.org/10.3969/j.issn.1674-621X.2025.04.015

Huang, Z., Li, G., & Zheng, Y. (2025). Empowering the high-quality development of science education through multimodal large models: Potentials, challenges, and applications. China Educational Technology, (6), 60–69. https://doi.org/10.3969/j.issn.1006-9860.2025.06.009

Jiang, H., & Sun, Y. (2025). Ethical reflections and governance strategies of multimodal large models in ideological and political education. School Party Building and Ideological Education, (6), 66–69. https://doi.org/10.19865/j.cnki.xxdj.2025.06.018

Wang, X., & Xu, X. (2024). Multimodal emotion recognition in online education considering credibility bias. Sensors and Microsystems, 43(11), 122–126. https://doi.org/10.13873/J.1000-9787(2024)11-0122-05

Wang, Y., Wang, Y. C., & Zheng, Y. (2021). Multimodal learning analytics: A new trend in intelligent education driven by multimodality. China Educational Technology, (3), 88–96. https://doi.org/10.3969/j.issn.1006-9860.2021.03.013

Yang, X., Bu, H., & Li, X. (2025). Promoting the deep application of multimodal large models in education: Value empowerment, scenario construction, and implementation strategies. Chinese Journal of Education, (4), 9–14.

Zhang, X., Li, W., Zhang, S., et al. (2023). Data-enabled teaching decision-making: From educational data applications to multimodal learning analytics for instructional support. E-educational Research, 44(3), 63–70. https://doi.org/10.13811/j.cnki.eer.2023.03.009

Author Biography

Yang Xiaoxue, Wuhan, Hubei, Wuhan Vocational College of Software and Engineering, Master, Associate Professor, Research area: Artificial Intelligence